Bacterial phase variation associated with repetitive DNA

Thesis

Bacterial phase variation associated with repetitive DNA.

Nigel John Saunders

B. Med . Biol. MB ChB. M.Sc. DipRCPath

A thesis submitted in partial fulfilment of the requirements for the Open University for the degree of Doctor of Philosophy

Discipline:

Medical Microbiology and Microbial Pathogenesis

December 1999

The Institute of Molecular Medicine

University of Oxford

Table of illustrations

Figures

Tables

Acknowledgements

I would like to express my sincere thanks to the many people who have made this period of research in Oxford as enjoyable, stimulating and productive as it has been. In particular I would like to thank:

Richard Moxon, for providing me with an opportunity and environment in which to spread my research wings and for his insightful and sage comments at many stages of this work.

Derek Hood, for his constant availability (when you know where to look) and advice on practical laboratory matters, his stoic attitude that things will eventually work, and his one liners.

John Peden, for sharing a vision of what could be done with whole genome analysis, doing all the computer programming and system manipulations, being a friend when little was working for either of us – and still being there when things are better.

Mike Gravenor, for being a mathemagician who was able to turn an (almost) infinitely long algebraic solution into a short easy to use formula, and his assistance in checking that I really had understood what all those maths papers were talking about.

Ian Peak, for advise on practical and non-practical aspects of lab survival, his example of how the apparently insurmountable can be overcome, and his continued friendship.

John Davies, for his openness and desire to share his long experience in neisserial research, his hospitality in his lab and home, and much red wine.

Alex Jeffries for his assistance with analysis of the MC58 genome sequence.

All the other members of the lab, past and present, who have helped me over the pasts few years.

And members of other labs who have helped me over the years, particularly Tara McDowell for help with gel shift assays.

Those who helped me in the past, by giving me opportunities and support to pursue my love of research and hence enabled me to make the best of this period with Richard, particularly: Jon Cohen, Douglas Burdon, Oleg Eremin, John Simpson and Paul Whiting.

Lori Snyder for her willingness to proof read (again and again), for her attention to detail, and her patience.

And last, but not least, the Wellcome Trust for the vision both to establish the Fellowships in Medical Microbiology and to give one to me!

Abstract

Phase variation is mechanism of phenotypic switching used by many pathogenic bacterial species. This thesis describes work on three aspects of phase variation. Mathematical models are described which can be used to determine the rate of phase variation and subsequently the influence of variation rate and fitness differences associated with the altered phenotype on population structure. An approach to whole genome analysis has been developed which has been used to identify putative phase variable contingency genes in *H. pylori*, *T. pallidum* and *N. meningitidis*. This has identified many new contingency genes likely to be involved in host - bacterium and bacterium population interactions. Finally, a detailed molecular investigation of the promoter of the phase variable *opc* gene of *N. meningitidis* is presented. In this it is shown that the promoter located homopolymeric tract controls transcription by affecting the relative spacing and facing of promoter components, that this determines RNA polymerase binding to the promoter, and that this interaction involves direct contact of the α-subunit of RNA polymerase with the promoter. In addition it is shown that transcription is dependent upon an IHF consensus sequence in the *opc* promoter.

Material that has been published:

The material in chapter 4 has been published in the paper:

Saunders, N.J., Peden, J.F., Hood, D.W., Moxon, E.R. (1998).

Simple sequence repeats in the *Helicobacter pylori* genome.

Mol Microbiol **27**: 1091-1098

Material from chapter 6 has been published in the paper:

Hervé Tettelin, Nigel J. Saunders, John Heidelberg, Alex C. Jeffries, Karen E. Nelson, Jonathan A. Eisen, Karen A. Ketchum, Derek W. Hood, John F. Peden, Robert J. Dodson, William C. Nelson, Michelle L. Gwinn, Robert DeBoy, Jeremy D. Peterson, Erin K. Hickey, Daniel H. Haft, Steven L. Salzberg, Owen White, Robert D. Fleischmann, Brian A. Dougherty, Tanya Mason, Anne Ciecko, Debbie S. Parksey, Eric Blair, Henry Cittone, Emily B. Clark, Matthew D. Cotton, Terry R. Utterback, Hoda Khouri, Haiying Qin, Jessica Vamathevan, John Gill, Vincenzo Scarlato, Vega Masignani, Mariagrazia Pizza, Guido Grandi, Li Sun, Hamilton O. Smith, Claire M. Fraser, E. Richard Moxon, Rino Rappuoli, and J. Craig Venter (2000)

The complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58.

Science (2000) **278**: 1809-1815.

And:

Nigel J Saunders, Alex C Jeffries, John F Peden, Derek W Hood, Herve Tettelin, Rino Rappouli and E Richard Moxon.

Repeat associated phase variable genes in the complete genome sequence of *Neisseria meningitidis*.

Molecular Microbiology (2000) **37**: 207-215.

Chapter 1

Bacterial phase variation in bacterial diversification and pathogenesis

1.1 The generation of diversity in bacterial populations

The generation of diversity within bacterial populations is important in the evolution and development of bacterial species and also in the adaptability of bacteria to their changing environments. Diversity is generated by a combination of programmed and random events that occur at different rates and confer different types of variability on the population. At one extreme there are random point mutations that occur throughout the coding and intergenic sequences that alter the expression, structure and function of bacterial components. At the other extreme, there are regulated responses that allow bacteria to control the expression of genes whenever the appropriate environmental conditions are encountered. Between these two extremes there is a variety of processes which adds to the capacity of a population to diversify, including: the presence and movement of insertion sequences that affect expression, mobile genetic elements that can move within and between populations, and the horizontal transfer of DNA between individual bacteria. One process that lies between the mutations that occur randomly throughout the genome and the programmed regulation of environmentally responsive genes is phase-variation. This process involves alterations in the cell at the level of DNA but in a way that generates predictable and predetermined adaptability for the bacterial population.

1.2 Definition of phase variation

Phase-variation describes a process of reversible switching between phenotypes, mediated by a genetic re-organisation, mutation or modification, which is not associated with a loss of coding potential, and which occurs at a comparatively high frequency. Phase-variation results in the continuous generation of alternative phenotypes within a population which facilitate adaptation to changing environmental conditions. Genes which undergo phase

variation have been called 'contingency genes' (Moxon *et al.*, 1994) which emphasises the evolutionary and functional implications of the variability of this subset of hypermutable genes. The switching of these genes occurs stochastically within the variable population.

1.3 Examples of bacteria that exhibit phase-variation

Phase variation has been recognised as a process associated with diversification for a long time (Andrewes, 1922). There are several examples, described in this chapter, that are familiar to many working in the field of infection, including the variation in H-antigens in *Salmonella* spp., of pili in *Escherichia coli* (*Esch. coli*), the major antigens of *Borrelia recurrentis* (BRE) in the relapsing fevers, and some surface proteins of *Neisseria meningitidis* (*N. meningitidis*). However, the range of bacterial species and the number of genes that display phase-variation are much broader than this and phase variation is recognised to be a common mechanism capable of generating enormous diversity within clonal populations of many bacterial species.

Phase-variation and phase-variable genes have been recognised in a wide variety of bacterial species, including Gram-positives, Gram-negatives, Spirochetes and Mollecutes. There are other frequent switching events that are not reversible (in the absence of horizontal transfer or further low frequency mutations) that cannot be considered to be phase-variation according to the previously stated definition. Sometimes the term phase variation has been used more loosely to describe any change (usually in colonial morphology) that occurs at a high frequency, e.g. one of the opacity variants in Group B Streptococci (Pincus *et al.*, 1992).

The phase varied genes and their potential roles in the processes relevant to bacterial virulence in the species in which these processes have been investigated in most depth will be described in the following sections.

1.3.1 Changes in *Borrelia* spp. surface proteins demonstrate immune evasion mediated by phase variation

One of the earliest and most dramatic consequences of phase variation during infection is illustrated in the *Borrelia* spp. Antigenic variations in *Borrelia* spp. involve the reversible selection of an expressed phenotype of its major surface proteins (VMPs) from a repertoire of possible genes within each strain. The consequences of this variation are most dramatically evident in the relapsing fevers. In relapsing fevers the infected individual experiences periods of fever that are interspaced by intervals of wellbeing. When the fever occurs large numbers of spirochetes can be found in blood smears and the borreliae disappear as the patient responds to them with specific antibodies. The waxing and waning of the borreliae populations is associated with the antigenic variation of the VMPs in the population within the host (Meleney, 1928). A studied strain of *B. hermsii*, an agent of relapsing fever, was found to express 26 antigenic variants in a mouse model with a rate of switching estimated to be 10^{-3} to 10^{-4} (Stoenner *et al.*, 1982; Barbour *et al*, 1982). VMPs differ in their molecular weights, peptide maps, and reactivies with serotype-specific antibodies (Barbour *et al.*, 1982 & 1983). Importantly, a serotype eliminated by neutralising antibodies from a first host may reappear in the populations in a non-immune second host (Meleney, 1928; Coffey & Eveland, 1967) – demonstrating the reversibility of the variation process.

1.3.2 Phase-variation in enteric bacteria illustrating variation of adhesion properties mediated by allelically stable structures

The earliest example of phase-variation described in the literature is of H-antigens in *Salmonella* spp. that were originally of interest in the development of typing schemes for the differentiation of bacterial strains (Andrewes, 1922). *Salmonella* spp. are typed on the basis of a serological scheme developed by Kauffman and White in which O-antigens

(based upon epitopes within LPS) and H-antigens (based upon epitopes on flagella) are identified using agglutination reactions. Not all salmonellae are able to alter their H-antigen. For example, *S. paratyphi A* and *S. abortusequi* are monophasic (in the absence of serological counter-selection) with phase I and phase II antigens respectively, and *S. gallinarium* is non-motile and thus has no H-antigens. Phase-variation of the O-antigen also occurs and was referred to as 'form-variation' to distinguish it from the flagellar variation (Kauffmann, 1954).

Phase variation in *S. typhimurium* (STY) converts the bacteria between adhesive and non-adhesive phenotypes based upon adhesion to porcine enterocytes (Isaacson & Kinsel, 1992). Cells of the adhesive phenotype differ from the non-adhesive cells in several ways including the production of fimbriae, more efficient uptake by phagocytes and longer intracellular survival, resistance to complement, and expression of 10 to 15 envelope associated proteins. The mechanism of phase-variation and the identity of the majority of the varied proteins are unknown. One of the co-regulated genes, other than the flagellum, is *rfaL*, an O-antigen ligase gene involved in the addition of O-antigen side-chains to LPS (Kwan & Isaacson, 1998).

Phase variation is also seen in the extended surface structures of *Esch. coli*. Type-1 fimbriae mediate mannose sensitive binding of *Esch. coli* to erythrocytes (Salit & Gotschlich, 1977), epithelial cells (Ofek & Beachey, 1978) and leukocytes (Bar-Shavit *et al.*, 1977). Expression of fimbriae is associated with the capacity to colonise mucosal surfaces (Silverblatt, 1974), but also increases susceptibility to phagocytosis (Silverblatt & Ofek, 1978). These structures are phase-variable (Brinton, 1959) and the adhesion that they mediate is considered to be the first step in the pathogenesis of urinary tract infections (Johnson, 1991). The altered expression and associated genomic rearrangements are seen in fresh clinical isolates (Abraham *et al.*, 1986) and the proportion of cells of each phenotype varies with the site of isolation (Lim *et al.*, 1998). One interpretation of this latter study is that cells bound to the uroepithelium are expressing the type-1 fimbriae,

whilst those in the urine are not, the surface bound population providing a continuous supply of fimbriae-negative variants to replace those lost through voiding. It is possible that a similar situation of 'resident' and 'mobile' populations also exists in the gut in which one subpopulation maintains stable colonisation and the other favors transmission. Interestingly, the rate of switching after isolation was higher in the isolates associated with more severe disease (acute pyelonephritis). The mechanism that controls the rate and direction of change of the fimbrial switch is environmentally responsive (see section 1.5.2), providing an example of a regulated stochastic process.

In addition to type-1 fimbriae, *Esch. coli* can also express another phase variable extended structure called pyelonephritis-associated pilus (Pap). Alternatively named the P adhesin, this structure binds to the P blood group antigen present on epithelial cells of the intestinal and urinary tracts. Phase variation of Pap is mediated by an entirely different mechanism to type-1 fimbriae (see section 1.5.6) but one of the factors which affects the expression of type-1 fimbriae, the leucine responsive regulator protein, also influences the switch controlling Pap. The multiple proteins that vary with H-antigens in *Salmonella* spp. and the overlap in the factors that influence switching in *Esch. coli* between differing phase-variable genes suggest that, at least in some species, there is considerable potential for the regulation and co-ordination of the stochastic processes that control phase variation.

There are several other examples of phase variation that are recognised in enteric bacteria, most of which have yet to be defined mechanistically. The Vi capsular antigen, composed of the polysaccharide galactosaminuronic acid, is present on some of the more virulent sub-species of *Salmonella* including *S. typhi*, *S. paratyphi C* and some strains of *S. dublin*. Vi is also present in some *Esch. coli* and *Citrobacter freundii* (*C. freundii*) strains where in the latter it is phase-variable (Johnson & Baron, 1969; Snellings *et al.*, 1981). *Serratia marcescens* displays a wide range of colonial morphologies after isolation from clinical specimens. At least one of these features, pigment production, is phase-variable (Bunting, 1940), and other bacterial components, such as flagellar components, also display features

of phase variability which may be related to changes in pigmentation (Paruchuri & Harshey, 1987). In addition, the mannose resistant flagella of *Proteus mirabilis* (Zhao *et al.*, 1997), and the acetylation of the K1 capsule of *Esch. coli* (Orskov *et al.*, 1979) - both virulence determinants of these important pathogens, are also phase variable.

Phase variation in enteric bacteria is not limited to the members of the enterobacteriaceae. Phase variation appears to provide a common strategy for adaptation in organisms that colonise the gastrointestinal tract. One example is the conversion between alginate and non-alginate production in *Pseudomonas aeruginosa* (Flynn & Ohman, 1988) – a major virulence determinant especially in the context of respiratory tract infections in patients with cystic fibrosis. Phase variation is also present in the vibrionaceae. *Vibrio cholerae* expresses several virulence factors including the toxin co-regulated pilus (Tcp) which is an adhesin the expression of which is regulated in response to similar environmental conditions to those that affect expression of the cholera toxin (Taylor *et al.*, 1987). Toxin production is under the control of a regulatory system that results in activation of toxin transcription through binding of the ToxR protein to the promoter region (Miller *et al.*, 1987). ToxR also activates the transcription of another transcriptional activator, ToxT, that in turn controls the production of Tcp (DiRita *et al.*, 1991). The capacity for ToxR to activate *toxT* is under the control of other factors including the product of *tcpH*, which is also a phase variable gene (Carroll *et al.*, 1997). This is an interesting example of the potential for the qualitatively different processes to be combined in the control of bacterial phenotypes.

1.3.3 Phase variation in *Bordetella* spp. illustrating that phase varied genes can also be subject to programmed regulation and can operate as part of co-ordinated systems and also phase variation of a transcriptional regulator

B. pertussis, the causative agent of whooping cough, has a number of identified virulence determinants, including:

Pertactin (per - pertussis, tactin - to touch) – a 69-kD protein that contains two Arg-Gly-Asp (RGD) adherence motifs where the amino-terminal motif promotes adherence to cell lines. Knockout of the gene reduces adherence of *B. pertussis* in *in vitro* models, and it elicits substantial immune responses acting as a protective antigen in animal models (Leininger *et al.*, 1991; De Magistris *et al.*, 1988; Shahin *et al.*, 1990).

Filamentous hemagglutinin (FHA) – a large filamentous protein containing an RGD adhesion motif that is both secreted and associated with the bacterial cell surface and mediates adherence to both ciliated and non-ciliated cells (Tuomanen & Weiss, 1985; Urisu *et al.*, 1986; Relman *et al.*, 1989; Cotter *et al.*, 1998). FHA stimulates an immune response in humans after clinical disease (De Magistris *et al.*, 1988) and protection can be synergistic with that produced by pertussis toxin (Robinson & Irons, 1983; Sato & Sato, 1984).

Pertussis toxin (PT) - a 105-kDa A-B toxin composed of five subunits (Locht & Keith, 1986; Nicosia *et al.*, 1986) which includes a surface associated adhesin that binds to cells through a lectin–like mechanism to carbohydrate receptors on the host cell surface (Brennan *et al.*, 1988; Tuomanen *et al.*, 1988) as well as having cilia specific, cell surface receptor and heparin binding properties (Locht *et al.*, 1993). The toxic subunit is an NAD-dependent ADP-ribosyltransferase which causes irreversible uncoupling of the regulatory GTP-binding proteins from their membrane receptors. The other subunits act as the targeting and delivery system. This affects several metabolic pathways and its effects include inhibition of adenylate cyclase (Hsia *et al.*, 1984; Katada & Ui, 1982) and

transducin (Manning *et al.*, 1984; Van Dop *et al.*, 1984). In addition, PT is a mitogen, an adjuvant, releases fatty acids from fat cells, interferes with chemotactic migration and alters vascular permeability (Hewlett *et al.*, 1983; Munoz *et al.*, 1981a & b). PT is immunogenic and antibodies against PT are protective. It also generates a haemolytic colonial phenotype on blood agar.

Phase-variation in *B. pertussis* affects the co-ordinated expression of pertactin, FHA, pertussis toxin (hemolysin - adenylate cyclase toxin), and also fimbriae (of which there are two serotypes). In *B. pertussis* strain Tohama, this is mediated by altered transcription of a gene with homology with two-component regulatory systems at the *vir* locus (Stibitz *et al.*, 1989). In this case programmed regulation and stochastic switching by phase-variation are closely integrated, demonstrating that phase-variation of a gene does not preclude the possibility of additional regulation. The expression of these factors is affected by growth conditions such as temperature and the concentration of $MgSO_4$. Under non-permissive conditions such as a temperature of 25°C or the presence of 20mM $MgSO_4$ these genes are repressed. When the cells are returned to permissive conditions then expression is resumed (Lacey, 1960; Idigbe *et al.*, 1981). These genes are under the control of the *bvg* (or *vir*) locus, which encode three proteins involved in sensory transduction (Stibitz *et al.*, 1988; Knapp & Mekalanos, 1988; Arico *et al.*, 1989; Weiss *et al.*, 1983). Therefore several of the virulence determinants of *B. pertussis* are part of a single regulon that is positively regulated by *bvg*, and the sensor / *bvg* system includes a gene which undergoes phase variation. When *bvg* is in the OFF state none of these genes are expressed, but when in the ON state they are expressed according to the environmental conditions.

However, this is not the full extent of the regulation of these phase-variable virulence genes. There is an additional regulatory factor that affects toxin expression but which has no effect on the adherence determinants (FHA and pertactin) (Carbonetti *et al.*, 1993). In addition, the fimbrial genes are independently phase-variable. The two serologically distinct fimbriae are composed of subunits of different molecular weight (Ashworth *et al.*,

1982; Irons *et al.*, 1985; Zhang *et al.*, 1985) and an individual strain can express both types, either type singly or have no fimbriae at all (when the *bvg* is in the ON state) – a process that can be observed to occur *in vivo* (Preston *et al.*, 1980). The *fim* genes in *B. pertussis* are phase-varied through alteration in the length of a promoter located homopolymeric tract of Cs (see section 1.3.5). The promoter of the *ptx* gene also contains a (shorter) homopolymeric tract at an equivalent location (Locht & Keith, 1986; Nicosia *et al.*, 1986; Nicosia and Rappuoli, 1987) which suggests that this might also be phase-varied independently. Taken together these mechanisms would provide a capacity to express at least 32 different phenotypic combinations of the components of the virulence regulon.

In addition to the altered expression of toxins and adhesins, *Bordetella* spp. also express variable LPS phenotypes which are influenced by environmental signals that are similar to those that influence the genes controlled by the *bvg* locus (Peppler, 1984; Peppler & Schrumpf, 1984; Caroff *et al.*, 1990; van den Akker, 1998). Alteration between the LPS phenotypes has been associated with altered susceptibility to antibacterial peptides (Banemann *et al.*, 1998), which form part of the defences present on mucosal surfaces and that act to control intracellular bacteria. This may contribute to the effects of phase variation on survival within phagocytes (Banemann & Gross, 1997). In some strains of *B. bronchiseptica* this phenotypic variation is under the control of the phase variable gene in the *bvg* locus. The mechanism of regulation and variation in other bordetellae is different and currently unknown (van den Akker, 1998).

The role of the phase variation in *Bordetella* spp. has not been established experimentally in infection. The expression of the adhesins is required for colonisation and therefore these must be present on isolates that make initial contact with cell surfaces and on those bacteria that are released to colonise new hosts. Asymptomatic culture-positive individuals can be detected during an outbreak (Krantz *et al.*, 1986) and *B. pertussis* is increasingly recognised to be an important cause of persistent cough in adults (Robertson *et al.*, 1987; Wright *et al.*, 1995; Deville *et al.*, 1995; Cattaneo *et al.*, 1996) - so not every host

colonised with *B. pertussis* develops whooping cough. It is possible that the altered phenotypes, by removing immunogenic structures from the cell surface, either delay clearance or facilitate persistence of a sub-population of bacteria (perhaps in an intracellular site) after the growth of the virulent variants is inhibited by the immune response. By these means the bacteria would be able to colonise an individual host for longer and prolong their infectious period.

Each of the phase variation events described above for which the mechanism has been determined involve slippage in homopolymeric tracts. However, the repertoire of genes and mechanisms used in bordetellae may be greater than are currently recognised. Hybridisation experiments indicate that *B. pertussis* possesses a homologue of the invertases of *Esch. coli* and *S. typhimurium* although the nature and role of this putative invertase are currently unknown.

1.3.4 Phase and antigenic variation in *Mycoplasma* spp. illustrating variation within families of similar proteins and the control of complex adhesion systems

Phase-variability and antigenic-variability are frequently combined to greatly increase the potential for surface diversification in a bacterial population. *Mycoplasma* spp. have no cell wall and no LPS containing outer membrane. Their single membrane presents a unique surface for interaction with their environment and host. As a group the mycoplasmas share a spectrum of phase and antigenically variable structures likely to play a role in both bacterially directed interactions with the host and immune evasion. The phase-variable systems in this group of bacteria reveal many features that are present in other bacteria that use phase-variation.

The first study of phase-variation in *Mycoplasma* reported high frequency, reversible changes in colony morphology, opacity and expression of a lipoprotein in *Myc. hyorrhinis* (Rosengarten and Wise, 1990). *Myc. hyorrhinis* is a species found in the respiratory tract of pigs in which it can cause rhinitis and arthritis. *Myc. hyorrhinis* surface proteins include a

group of 3 lipid modified proteins, VlpA, VlpB and VlpC, which undergo phase-variation and size variation due to duplication, deletion and recombination within a highly repetitive C-terminal region (Rosengarten & Wise, 1990 & 1991, Yogev *et al.*, 1991). Vlps have been shown to be targets for immune damage to mycoplasmas, to be involved in interactions with host cells (Rosengarten and Wise, 1991) and both mechanisms of variation generate escape variants from growth-inhibiting antibodies (Citti *et al.*, 1997). The combinatorial effect of size and phase-variation has been estimated to be able to generate over 10^4 structural permutations of Vlps. Further, that estimate was made prior to the demonstration that individual strains could have 6 or 7 *vlp* genes (Yogev *et al.*, 1995) and does not consider the possibility of recombination between them.

Mycoplasma bovis (*Myc. bovis*) is a species that usually exists as a harmless commensal in the respiratory tract, and can also be isolated from the milk of healthy cattle. It also causes bovine mastitis, arthritis, pneumonia, subcutaneous abscesses, meningitis and infertility. *Myc. bovis* has the capacity to express a repertoire of variable surface expressed lipoproteins called Vsps. These are different from the Vlps described above (Behrens *et al.*, 1994). This variability includes both size and antigenic variation, due to alterations in the number of repetitive structural units of different sizes located predominantly at the exposed C-terminal end of the proteins, as well as phase-variation of expression (Rosengarten *et al.*, 1994). Infection studies demonstrated that there is substantial variation in the surface antigens expressed in serial isolates from single animals, and *Myc. bovis*-infected cattle develop strong preferential antibody reactions to these proteins (Rosengarten *et al.*, 1994) demonstrating that the variability that can be detected *in vitro* occurs during natural infection and that this is potentially a means of immune evasion. *In vitro* the phase-variation of each Vsp was independent. In addition to the capacity to vary the expression of Vsps, *Myc. bovis* colony immunostaining exhibits additional phase-variation that is not due to size variation or to altered expression of the proteins as detected by Western blotting – suggesting the presence of a phase-variable process that can mask Vsp epitopes (Behrens

et al., 1994). The Vsps of *Myc. bovis* and the Vlps of *Myc. hyorrhinis* share several features: i. they are both anchored to the surface as lipoproteins, ii. there are some shared epitopes, iii. they both have a surface exposed C-terminal region with an extensive and size variable structure, iv. they exhibit independent phase-variation with the capacity for combinatorial expression, v. they are the major antigens on the membrane surface. However, there are also differences: i. carboxypeptidase digestion reveals a digest pattern that shows a regular pattern in Vlps (suggesting tandem repeats of very similar subunits) and an irregular pattern in Vsps (suggesting sets of multiple sets of non-similar repeats), ii. Vsps are much more resistant to trypsin digestion, iii. Vsps are able to alter their size by variations in sites other than the C-terminal end. It is not possible to determine, on the basis of the available evidence, whether this represents great divergence in a common ancestral system or whether it represents an example of convergent evolution within two *Mycoplasma* spp. Either alternative suggests that phase variation is important in the generation of diversity in *Mycoplasma* spp. and in host interactions.

Vsps are not the only phase-variable proteins identified in *Myc. bovis*. pMB67 is a 67 kDa protein that is not lipid modified and does not contain a Vsp-like repetitive periodic C-terminal structure but which is size variable (although at a lower frequency than Vsps), phase-variable and immunogenic (Behrens *et al.*, 1996). Only one of the proteins identified by Western blotting with the antibody reactive with pMB67 has been characterized but individual strains have been shown to express up to three reactive proteins and there may be five potentially phase-variable proteins related to pMB67 in individual strains of *Myc. bovis*.

Further examples of the evolution of phase-variable structures with related properties are present in *Myc. fermentans*, a putative human pathogen, and *Myc. gallisepticum*, a respiratory pathogen of poultry. *Myc. fermentans* has a family of at least 7 phase-variable surface lipoproteins and also varies other membrane components (Wise *et al.*, 1993; Theiss *et al.*, 1993). One of these, P78, has been shown to be expressed as part of an operon

believed to encode an ABC transporter system within which it is likely to be a substrate binding protein (Theiss & Wise, 1997). Using an antibody reactive with VspA from *Myc. bovis*, a size- and phase-variable non-lipoprotein (PvpA) was identified in *Myc. gallisepticum* that is expressed in addition to three variable lipoproteins similar to Vlps or Vsps. PvpA does not have homology with VspA, as determined by DNA hybridisation, and there is only one copy in the *Myc. gallisepticum* genome – whilst there may be as many as 7 Vsp homologous loci present (Yogev *et al.*, 1994). It is not known whether PvpA and pMB67 are similar proteins.

Many mycoplasmas have a flask-like shape with a tip at one of the poles. This 'tip organelle' or 'terminal structure' functions in host cell contact and attachment. The ability to adhere is correlated with the capacity to bind sialylated receptors on epithelial and red blood cells in both *Myc. pneumoniae* and *Myc. gallisepticum*. In *Myc. pneumoniae* two surface proteins: P1 (Hu *et al.*, 1992) and a 30kDa protein, P30, (Baseman *et al.*, 1987) located at the tip organelle make the initial contact with epithelial cells. Spontaneous non-hemadsorbing phase-variants that do not express P30 and do not localize P1 to the tip organelle are non-adhering and avirulent. Revertants that re-synthesize P30 express both P1 and P30 at the tip organelle and regain cytadherence and virulence capacity (Krause *et al.*, 1982 & 1983; Baseman *et al.*, 1987). Antibodies generated to the P30 protein localize to the tip organelle and block adherence (Morrison-Plummer *et al.*, 1986). Phase variation of adherence involves more proteins than just P1 and P30. The most prevalent class of cytadherence-negative variants lack 4 high molecular weight proteins that phase vary simultaneously (Krause *et al.*, 1983). The mechanism for this coordinated phase variation is unknown. Cloning of one of the proteins, HMW3, and investigation of variants did not reveal any genome rearrangements associated with variation (Ogle *et al.*, 1991). Analysis of the genome sequence of *Myc. pneumoniae* (Himmelreich *et al.*, 1996) reveals that the 6 recognised adhesion proteins associated with hemadsorption are located in 4 loci: P1 is located apparently operonically with its associated protein 'ORF6', P30 is located with

HMW3, and HMW1 and HMW2 are located in 2 separate locations (personal observation). This suggests that the expression of HMW3 and ORF6 are dependent upon that of the associated ORFs and that variation of a regulator that acts on the other loci is likely to mediate the observed phase variation. In *Myc. gallisepticum* protein profiles were studied in hemadsorption phase-variants and revealed a similar situation. Altered binding is not associated with the altered expression of only one protein but with 4 or 5 which appear to be expressed phase-variably in a coordinated fashion (Athamna *et al.*, 1997). This study did not address whether the phase-variation of these proteins is independent or coordinated but it does suggest that there is coordinated expression of these proteins and thus that there may be a variable regulator protein present within *Myc. gallisepticum*.

Myc. pulmonis, a respiratory pathogen of mice, varies one of its major surface antigens, V-1, encoded by the *vsa* genes, with a combination of antigenic and phase-variation. Unlike *vsp* and *vlp* gene expression, in which the switching of each variant is independent, *Myc. pulmonis* expresses only a single version of this protein at a time. The studied strain had 7 expressible variant ORFs and in addition each contains a series of repeats that are unstable making them capable of generating length variants of these proteins (Bhugra *et al.*, 1995). Both the mechanism of variation and the nature and location of the repeats differ from the other surface proteins described in this section (see section 1.3.2) but the nature of the resultant phenotypic variation is similar. *Mycoplasma hominis* also has a variable surface protein with repetitive motifs that can generate size and antigenic variation which is phase-varied by a different mechanism again, variation in a homopolymeric tract of As (Zhang & Wise, 1996 and 1997) (section 1.5.4).

Taken together, the observations from the various *Mycoplasma* spp. reveal that they have an enormous capacity to generate structurally, antigenically and functionally versatile surfaces. The variety and variability of these surface proteins enables these species to exist as mixed populations with potentially different properties that may be important at different stages of colonisation and infection. This may partly explain why they are

associated with such a wide spectrum of diverse and chronic infections. Taken together the phase-variable systems of mycoplasmas (many of which are also antigenically variable) allow a single strain, of a bacterial species with the smallest genomes of any free-living bacteria, to express many thousands of variant phenotypes. Between the various species studied a variety of surface proteins, some of which are adhesins, transport systems, restriction-modification systems and perhaps a transcriptional regulator have been shown to be phase-variable. However, what is missing to date for the mycoplasmas is a complete picture of the variable repertoire in a single strain or species.

1.3.5 Phase variation in *Haemophilus influenzae* describing the recognition of the association between simple sequence repeats and phase variation and their use to identify phase variable genes. Describing the first attempt to determine the whole repertoire of phase variable genes, and the complex coding capacity conferred by variation of groups of genes affecting distinct but related functions.

Haemophilus influenzae (*H. influenzae*) is a common cause of bacterial meningitis in unvaccinated populations, respiratory tract infections and otitis media. The ability of *H. influenzae* to display variations, as indicated by changes in colonial morphology, was recognised in early reports (Pitman, 1931). A substantial amount of information on phase-variable genes in *H. influenzae* was available prior to the publication of the complete genome sequence (Fleischmann *et al.*, 1995) which provided a means to define the complete repertoire of repeat associated phase variable genes in a single strain. The study of phase variable genes in *H. influenzae* illustrates the development of repeat-based searching methodology in this field.

1.3.5.1 Phase variation of fimbriae

The first example of phase variation recognised in *H. influenzae* is mediated by one of the most elegant examples of regulation controlled by promoter located repeats (see section 1.5.5). Several early observations reflect the phase variation of this structure. One virulence determinant of *H. influenzae* is the LKP family of fimbriae which mediate adherence to mucosal cells (Brinton *et al.*, 1989) and mucus (Read *et al.*, 1991). During natural infection, nasopharyngeal isolates are often fimbriated whilst isogenic isolates from systemic sites are usually non-fimbriated (Pichichero *et al.*, 1982; Guerina *et al.*, 1982; Mason *et al.*, 1985). Piliated phase variants can be obtained from most non-piliated isolates by selective hemadsorption, making use of the haemagglutination properties of the pili (Connor & Loeb, 1983; LiPuma & Gilsdorf, 1988). *In vitro* cell adhesion studies of *H. influenzae* to buccal epithelial cells show that, after 30h of incubation, cells incubated with non-piliated *H. influenzae* become coated with piliated variants (Patrick *et al.*, 1989). These observations suggest that fimbriae are selected for during colonisation of the nasopharynx but that they are lost during the steps that lead to invasive disease. This is supported by observations of phase variation and interactions of piliated and non-piliated *H. influenzae* with human cells (Farley *et al.*, 1990). *hifA* encodes the major fimbrial subunit and *hifB* encodes the fimbrial chaperone and they are divergently transcribed from a common overlapping promoter. The genes 3′ of *hifB* encode the other genes needed for expression of fimbriae (reviewed in Gilsdorf *et al.*, 1997). The *hifA* and *hifB* genes are simultaneously phase varied by changes in this promoter region (van Ham *et al.*, 1993). Thus this switch simultaneously controls both the export system and the production of the fimbrial components. In the context of the role of fimbriae in virulence it is interesting to note that the hypervirulent acapsulate invasive *H. influenzae* biogroup aegyptius strains (the cause of Brazilian purpuric fever) do not lose their pili on subculture in the same way as other *H. influenzae* strains (Davis *et al.*, 1950; Read *et al.*, 1996b).

1.3.5.2 Phase variation of LPS

Lipopolysaccharide (LPS) is the predominant constituent of the outer membrane of Gram-negative bacteria (Nikaido, 1996) and is therefore one of the bacterial structures that is available to mediate interactions between bacterium and host. *H. influenzae* has an LPS that lacks the long O-side chain that comprises the O-antigen of most enterobacteriaceae. Therefore it does not have the capacity to mask its core LPS structures by expressing a variety of O-side chains, as seen for example in *Salmonella* spp. Instead it is the sugars and substitutions of the LPS core that are available to interact with the host. There is variability in these structures between strains and also within a single strain and there is both co-ordinate and independent phase variation of multiple LPS epitopes as indicated by binding to anti-LPS monoclonal antibodies (Kimura & Hansen, 1986; Weiser *et al.*, 1989a). The study of the function of individual phenotypes is complicated by the extent of the LPS variability. However, there is some evidence that particular phenotypes affect survival *in vivo* (Kimura & Hansen, 1986) and susceptibility to serum killing (Gilsdorf *et al.*, 1986) and that isolates from the nasopharynx and systemic sites often differ in LPS phenotype (Mertsola *et al.*, 1991).

The locus responsible for the expression of variable expression of epitopes detected by three different antibodies, *lic*1, was identified and both knock-out experiments and transformations with *lic1* of a strain which lacked the associated LPS epitopes confirmed its role in LPS biosynthesis (Weiser *et al.*, 1989a). The mechanism mediating phase variation was sought in this locus which includes four genes, the first of which, *lic1A*, was found to contain a (CAAT)n tetrameric repeat at the 5′ end of the coding sequence – alterations in the length of which alters the translational reading frame of the sequence 3′ of the repeat (Weiser *et al.*, 1989b). These investigations explained only 2 of the then recognised 5 variable LPS epitopes (in strain RM7004) so these were pursued by probing the genome with the repeats identified in *lic1A* (Weiser *et al.*, 1990). (CAAT)n repeats were identified in two other genomic loci that were called *lic*2 and *lic*3 in RM7004 and

similar probing in other strains revealed from 2 to 5 potentially variable loci based upon the presence of the (CAAT)n repeat. Site-directed mutagenesis of the repeat associated ORF in *lic2* abolished phase variation of this gene and identified sequences required for the expression of additional LPS epitopes (Szabo *et al.*, 1992).

A (CAAT)n associated LPS biosynthetic gene, *lex*-1, was identified through a library screen search for genes that restored the wild-type LPS phenotype and virulence to mutants previously found to have a reduced ability to cause invasive disease in the infant rat model (Cope *et al.*, 1990; Cope *et al.*, 1991). This sequence was derived from a strain with 4 (CAAT)n associated loci – as indicated by Southern hybridisation. It was concluded that because of a difference in the *XbaI* digest pattern of the cloned fragment and divergence from the limited sequence associated with the *lic2* locus from its original description that this was not *lic2*. The subsequent description of *lic2A* (High *et al.*, 1993) stated that 'A gene with similar features has also been described termed *lex-1*' which suggests that whilst similar they are not the same. In fact, the Genbank submitted sequence of *lic2A* (which lacks the first 78 bases in the paper) is identical to bases 159 to 1055 (of 1860) of *lex-1* apart from the number of CAAT repeats present and a single base polymorphism at *lex-1*$_{967}$. The divergent *XbaI* site is not the one in the sequenced region. *lex-1* and *lic2A* are the same gene.

The *lic3* locus was cloned and found to be composed of four closely apposed open reading frames (Maskell *et al.*, 1991). The first ORF contained the repeats and the second the UDP-galactose-4-epimerase gene (*galE*) – which is known to be involved in LPS biosynthesis. A deletion/insertion mutant extending from the 3′ portion of the first ORF to the 5′ portion of the third ORF had an altered LPS phenotype and reduced virulence in the infant rat model of *H. influenzae* infection following both intranasal and intraperitoneal inoculation.

The same group that described *lex-1* also identified another locus, *lex-2*, using a similar approach which also affected the expression of similar surface structures (epitopes reactive

with 5G8 and 4C4 monoclonal antibodies). The first gene in this locus, *lex-2A*, contained 18 tandem repeats of the motif GCAA at its 5′ end (Jarosik & Hansen, 1994). Similar phase variation in the LPS of *Haemophilus somnus*, (a pathogen of cattle) had been described (Inzana *et al.*, 1992). Subsequently, an LPS biosynthetic gene that has moderate homology (48% amino acid homology) with the second gene in the *lex-2* locus, *lex-2B*, was identified which has (CAAT)n repeats which affect translation in *H. somnus* (Inzana *et al.*, 1997).

1.3.5.3 The use of repeats to identify phase variable genes

Based upon the association of LPS genes with tetrameric repeats and the repeat mediated variation in fimbrial expression it was hypothesised that these repeats could be used as markers with which to search the completed genome sequence of *H. influenzae* strain Rd (Fleischmann *et al.*, 1995). In essence this is an *in silico* means of performing the same experiments that were performed using repeat probes in Southern blotting. Using a combination of BLASTN and FINDPATTERNS search algorithms, repeats have been sought as markers for phase variable genes. Using this method a list of repeat associated phase variable genes was compiled providing the first example of a bacterium in which the complete repertoire of phase-variable genes in a single strain was thought to be known. The presence of each repeat in the genome was sought, one at a time, using a probe sequence. In the analysis of *H. influenzae* strain Rd, this was done for each of the 256 possible tetranucleotides and for each of the repeats composed of shorter motifs (homopolymeric tracts, dinucleotide- and trinucleotide repeats) (Hood *et al.*, 1996). A second similar analysis, using a different search algorithm, generated similar but in some instances apparently discordant results (van Belkum *et al.*, 1997a). These discrepancies have been resolved by performing a repeat analysis of the *H. influenzae* strain Rd genome using the ACEDB system (described in Appendix 1).

The search by Hood *et al.* identified a number of novel repeat associated genes (see table 9.1 in Appendix 1) for which variation in the number of repeats between unrelated strains could be demonstrated – suggesting that they are unstable and mediate phase variation. These novel genes included 4 related iron binding proteins, each with similar repeats. These may be important in immune evasion and/or acquisition of iron from different carrier molecules. A novel LPS biosynthesis gene, *lgtC*, was identified which although not intact in *H. influenzae* strain Rd was present, functional and demonstrated to be involved in the synthesis of epitopes related to those altered by *lic2* and *lex2* in other strains. In addition an adhesin, a methyltransferase, and two genes of unknown function on the basis of homologies were identified.

In addition to the search for repeat associated potentially phase-variable genes, the availability of the complete genome sequence also facilitated the investigation of the LPS biosynthetic pathways in this strain of *H. influenzae* (Hood *et al.*, 1996). Combining the information from this study and those of the phase variable LPS genes facilitates an interpretation of the phase variable repertoire of LPS structures in *H. influenzae*.

In addition to the tetramer associated genes Hood *et al.*, also noted a trimer in the promoter region of *hxuC*. This is an outer membrane protein that is required for *H. influenzae* to utilise free heme at low concentrations (Cope *et al.*, 1995). The previously sequenced version of *hxuC* from *H. influenzae* strain DL42 had a different number of repeats from those seen in *H. influenzae* strain Rd.

The search by van Belkum *et al.* (1997a) reported the repeats previously found by Hood *et al.* and five additional repeats with pentamer and hexamer motifs (one of which was the pentameric repeat thought to be located at the origin of replication (Fleischmann *et al.*, 1995)) that they suggested might mediate phase variation. In addition they also identified 5 dinucleotide repeats with 5 copies of the component motifs but did not suggest whether they believed that they were involved in phase variation. This paper therefore identified 4 previously unreported repeat loci that might mediate phase variation of associated genes.

Of these, 3 were hexameric repeats. Two of these hexameric repeats are located within open reading frames and, although they might alter antigenic or functional properties of the encoded proteins, would not be expected to alter gene expression. The third is located 5′ of a reading frame in a location such that it would not be expected to affect promoter function (see Appendix 1).

1.3.5.4 Limitations of the method using repeat to identify phase variable genes

This approach is very powerful but these analyses of *H. influenzae* also demonstrate its limitations. It is important to recognise that *H. influenzae* as a species has not been sequenced – *H. influenzae* strain Rd has. The analysis of this type of representative strain does not provide an opportunity to determine the whole spectrum of phase variability within a species – although it does provide a useful reference point from which to investigate the differences between strains. Firstly, a search is limited by its search parameters. For example, a repeat associated methyltransferase was not identified in the first reported analysis of the genome because it is associated with a pentameric repeat and the original search focussed upon repeats with motifs of 1 to 4 bases in length. Secondly, not all phase variable genes present in the species are necessarily present in a single representative strain: Southern hybridisation experiments suggest that there are at least two loci associated with (CAAT)n repeats and two more with (GCAA)n (Jarosik & Hansen, 1994) that remain to be identified in other strains of *H. influenzae*. Thirdly, there may be other processes mediating phase variation in *H. influenzae* that have yet to be identified in this species which may, for example, include the mechanism which underlies reversible phase variation of the *H. influenzae* capsule which has not yet been determined.

Finally, analysis of the *H. influenzae* strain Rd genome has not resolved the picture with respect to one type of phase variation in *H. influenzae* colonial morphology. Phase variation in colonial appearance in *H. influenzae* between opaque (O), intermediate (I) and transparent (T) phenotypes has been described that is associated with changes in the LPS

but does not correlate with the variation in antibody reactivity associated with changes in the *lic* loci (Weiser, 1993). O and T phenotypes differ in their virulence in the infant rat model and in their ability to colonise the rat nasopharynx. A library prepared from an O variant of *H. influenzae* strain Rd was screened by transforming T phenotype bacteria for O phenotype transformants. This identified two genes *oapA* and *oapB* (Weiser *et al.*, 1995). Although transformation with chromosomal DNA from the O variant gave O transformants this could not be repeated with the prepared library. Instead, a different I phenotype (to which *H. influenzae* strain Rd did not normally vary) was identified and this clone was studied. There is no indication that this phenotype is the same as the I phenotype previously described in other strains. Furthermore, a knock out of *oapA* in a strain (Eagan) which varies at high frequency between O, I and T lost the ability to express a T phenotype but retained O to/from I switching, even though the original screening experiment was constructed to identify the O associated gene. A fusion of *oapA* with the reporter gene *phoA* in a transformant of strain Eagan showed high frequency phase variation in the secretion and production of this gene. One interpretation of these results is that the three phenotypes (O, I and T) are controlled by at least two phase variable genes, *oapA* affecting the T phenotype, and an unrelated gene (which affects LPS biosynthesis or substitution) generating the O phenotype. Identification of the T associated gene using an O variant derived library would then be entirely fortuitous due to heterogeneity in the strain, or due to transformation with a second locus variant that is required for the expression of the O phenotype from which the library was prepared. In a study that did not address the I phenotype, it was concluded that that O ↔ T phenotypic variation is due to changes in the amount of cell associated capsular polysaccharide and that this transition is associated with changes in the underlying LPS (Roche & Moxon, 1995). This was extrapolated, following the description of *oapA*, to suggest that the T ↔ I transitions and phenotypes could only be discerned in non-O cells (i.e. those with less cell associated capsule) (Moxon *et al.*, 1996). This explanation still leaves some issues unresolved. First, the original description

included O, I and T variants of the unencapsulated strain Rd-T (Weiser, 1993) which raises the possibility of another variable phenotype. Second, the reported LPS changes seen in association with the O phenotype are less complete in the report of Roche and Moxon (1996) than originally reported. Finally, the mechanism of phase variation for both phenomena remains unknown. It was suggested that the *oapA* switch is controlled by a mechanism that is not mediated by alterations in repeat tract lengths (Weiser *et al.*, 1995) – which is the only mechanism mediating phase variation recognised to date in *H. influenzae*. However, since the only strain for which corresponding sequence information is reported is *H. influenzae* strain Rd, which was selected for the study because it does not phase vary this phenotype at high frequency, this conclusion is premature.

1.3.5.5 An example of the investigation of phase variation in whole animal systems

The gene containing the repeats in *lic*-1 has been investigated further and it does not encode a glycosyltransferase or other gene involved in the manufacture of LPS sugar components. It has homology with eukaryotic choline kinases and *H. influenzae* has the capacity to link choline acquired from the environment to its LPS in a phase variable fashion (Weiser *et al.*, 1997; Risburg *et al.*, 1997; Schweda *et al.*, 1997). This gene is now thought to act in the substitution of LPS with choline. Strains expressing the phosphorylcholine (ChoP) epitope are more sensitive to serum killing involving C-reactive protein (CRP) (Weiser *et al.*, 1997; Weiser *et al.*, 1998a). There is a bias towards expression of ChoP in human isolates and there is evidence that suggests that the presence of ChoP contributes to persistence in the respiratory tract in animal models (Weiser *et al.*, 1998a). It has subsequently been proposed that variation between LPS which carries the structure determined by *lic2* (which confers some resistance to antibody mediated serum bactericidal activity) and the ChoP substituted LPS mediated by *lic1* (with resistance to CRP) adapts *H. influenzae* to different host microenvironments (Weiser & Pan, 1998). The contribution of ChoP to survival in the host may be more complex than solely its effects on

CRP dependent killing. For example ChoP is known, in other contexts, to affect persistence, invasiveness and to influence lymphocyte responsiveness and cytokine production (reviewed in Harnett & Harnett, 1999).

1.3.6 Phase variation in *Neisseria* spp. and particularly *Neisseria meningitidis*, illustrating the extent to which phase variation determines the expression of bacterial structures which interact with the host.

Neisseria meningitidis is one of the most intensively studied organisms that exhibit phase variation. *N. meningitidis* is a Gram-negative diplococcus, is a common commensal species colonising the upper respiratory tract, has the capacity to cause septicaemia, and is the principal cause of bacterial meningitis. It is the severity of these infections and the lack of an effective vaccine against strains with the serogroup B capsule that has sustained a high level of research into this organism. The focus of much of the research has been directed towards the components of the bacterial surface and many of these are phase variable. In addition *Neisseria gonorrhoeae*, the cause of gonorrhoea, is a closely related organism which is safer to handle in the laboratory – information from which can be used to complement information derived from *N. meningitidis* with which it shares many virulence determinants.

1.3.6.1 The *N. meningitidis* capsule

The bacterial capsule of *N. meningitidis* is a major determinant of bacterial survival once organisms are disseminated in the blood stream (DeVoe, 1982) and is the basis of the division of the species into 12 serogroups (Jennings *et al.*, 1977). Only a few of these serogroups are associated with disease. Serogroup A causes the majority of epidemic-associated cases in sub-Saharan Africa whilst the majority of disease in the northern hemisphere is caused by serogroups B and C. Serogroup B and C capsules are composed

of α2-8 and α2-9 linkages respectively of polysialic acid. The importance of these structures in association with invasive disease and meningitis is emphasised by the observation that other bacteria which cause meningitis, *Esch. coli* and group B streptococci, have a similar capsule to serogroup B meningococci (Kasper *et al.*, 1973 & 1983). Several of the capsular structures are immunogenic and form the basis of meningococcal vaccines. The structure of the group B capsule is present on N-CAM which is present within the human brain (Finne *et al.*, 1983) and in other sites (Finne *et al.*, 1987). Therefore, it is not recognised as a foreign epitope and is not immunogenic in man, and thus is not a useful vaccine candidate. When antibodies are formed against this structure, as seen in rare cases of myeloma, they do confer protection by adoptive transfer in murine infection models and also bind to calf brain cells (Azmi *et al.*, 1995). A vaccine is still needed that is effective against serogroup B meningococci and in this context it is important to understand the structure, function and behaviour of the meningococcal cell surface components.

The expression of the serogroup B capsule is phase variable (Hammerschmidt *et al.*, 1996a & b). The capsular structure is an accepted virulence determinant of *N. meningitidis* but its function in the normal transmission – colonisation life cycle of the bacterium is unknown. Not all capsular serotypes are associated with infection and the extent to which the others display phase variation has not been defined. The selective pressure for the presence of capsule in the avirulent serotypes must be unrelated to the development of invasive disease. Further, since invasion is a very rare event and almost all invasive isolates must be descended from progenitors that have never caused invasive disease this must be true for the invasive isolates too. One possible function of the capsule is as an anti-desiccant for organisms in the environment. In contrast to the universal presence of capsule in blood stream isolates a significant proportion (50% in one study) of colonising bacteria are acapsulate (Cartwright *et al.*, 1987) and the remainder tend to express far less capsular polysaccharide than case isolates (Mackinnon *et al.*, 1993). Loss of capsule has been

recognised to be associated with facilitating the adhesion to cell surfaces by other surface proteins (see below) and to prevent adhesion and invasion of *N. meningitidis* into epithelial and endothelial cells (Virji *et al.*, 1992a & 1993b; Stephens *et al.*, 1993; Hammerschmidt *et al.*, 1996a). An alternative possible selected function for capsule is actually the reverse of this process: that it acts as a releasing mechanism such that phase-variation from OFF to ON prevents stable binding to the cell surface. This may result in the continuous generation of a transmissible, perhaps also environmentally adapted, population from a resident and relatively stable colonising population. It should be noted that in the mouse intranasal model of infection the capsule has been found to be required for colonisation (Mackinnon *et al.*, 1993) - the interpretation of which is unclear.

1.3.6.2 Phase variation of pili, pilus associated proteins and pilus modification

Bacterial cell surface sialic acids interfere with the immune system through the alternative complement activation pathway (Fearon, 1978; Jarvis, 1995) which is required to respond to *N. meningitidis* (Nicholson & Lepow, 1979) and the genes in the capsule locus confer serum resistance upon meningococci (Hammerschmidt *et al.*, 1994). The presence of capsule also reduces adherence and uptake into macrophages and delays or prevents the killing of phagocytosed bacteria (McNeil *et al.*, 1994; Read *et al.*, 1996a). In this way the capsule is thought to provide an immune evasion mechanism in invasive disease isolates.

It is thought that the pilus is the only surface structure that is extended sufficiently to mediate initial contact and adhesion to mucosal cells in encapsulated bacteria. The first indication of their association with virulence came from the demonstration that only colony types of *N. gonorrhoeae* (an acapsulate species) that are piliated are virulent in human volunteers (Lambden *et al.*, 1979). *In vitro*, expression of pili is required for adhesion of encapsulated bacteria to epithelial and endothelial cells (Stephens & McGee, 1981; Virji *et al.*, 1991; Nassif *et al.*, 1994), is associated with increased adherence to mucosal cells and a tropism for nonciliated cells with microvilli (Rayner *et al.*, 1995) but

has no affect on monocyte interactions (McNeil & Virji, 1997). In cells with L3 sialylated LPS, pili are required for adhesion in both capsulate and acapsulate bacteria and increase epithelial and endothelial cell invasion (Virji *et al.*, 1995). In addition pili are associated with twitching motility, competence for natural transformation (Rudel *et al.*, 1995c) and bacterial autoagglutination. Whilst pilin, as the structural component of pili, is required for pilus-mediated adhesion it has not been conclusively shown to act as an adhesin. However, adhesion does vary between strains with different variant pilins (Lambden *et al.*, 1980; Virji *et al.*, 1992b & 1993b; Nassif *et al.*, 1993). The basis for this variability has not been resolved but the ability to facilitate adhesion is linked to the formation of bundled pili (Marceau *et al.*, 1995). The ability to form bundled pili may be influenced by charge and other differences between pilin variants and also by modification of the pili by sugars. To date, two sugars: an O-linked galactose α1,3 N-acetyl-D-glucosamine (GlcNAc), and digalactosyl 2,4-diacetamido-2,4,6-trideoxyheptose have been demonstrated to be linked to pilins (Stimson *et al.*, 1995; Parge *et al.*, 1995; Marceau *et al.*, 1998). Changes in pilus glycosylation has been linked to altered adherence (Virji *et al.*, 1993b) and loss of glycosylation has been shown to favour agglutination of pili and the formation of bundles but the role of glycosylation in adherence and pathogenesis is disputed (Marceau *et al.*, 1998). Further modifications of pilin by glycerophosphate and phopshorylcholine have also been described (Stimson *et al.*, 1996; Weiser *et al.*, 1998b) the functions of which are currently unknown. In addition to pilin, another protein, PilC, is essential for pilus mediated adhesion and for pilus biogenesis (Rudel *et al.*, 1992). PilC was first described as a protein that co-purified with the pilus (Jonsson *et al.*, 1991), has been shown to be associated with both the tip of the pilus (Rudel *et al.*, 1995a) and the outer membrane at the pilus base (Rahman *et al.*, 1997) and to compete with pilus mediated adhesion (Rudel *et al.*, 1995a). These locations may reflect the dual functions of the protein in pilus-mediated adhesion and assembly respectively. Most strains possess two variant *pilC* genes that encode proteins, PilC1 and PilC2, that share approximately 70% sequence identity

although some strains lack *pilC2* (Jonsson *et al.*, 1991; Nassif *et al.*, 1994; Pron *et al.*, 1997). Either PilC protein can function in pilus assembly whilst only PilC1 expression confers adherence when expressed. It is therefore thought that PilC1 contains the cell binding domain(s) (Nassif *et al.*, 1994; Rudel *et al.*, 1995a). In addition, the promoter regions of these two genes are divergent and they respond differently to environmental conditions. PilC2 is not affected by the interaction with host cells whilst PilC1 expression is increased during initial contact (Taha *et al.*, 1998), which is, at least in part, controlled by PilA (Taha *et al.*, 1996) - a regulatory protein that, as demonstrated in *N. gonorrhoeae*, has additional effects which include increasing serum resistance (Taha, 1993). These observations are consistent with the proposed roles of these proteins in pilus assembly and adhesion. The situation in *N. gonorrhoeae* is broadly similar but differs in that both PilC1 and PilC2 confer adherence properties (Nassif *et al.*, 1994).

N. meningitidis vary the structure, function and expression of pili by a number of mechanisms. As described above, the expression of *pilC1* is regulated in response to environmental conditions. In addition, stochastic processes affect both the sequence of the expressed pilus, its substitutions and pilus expression. The sequence of the expressed pili is altered by non-reciprocal exchange of the expressed reading frame between promoterless silent copies (*pilS* genes) and a single expression site (*pilE*) (see section 1.5.1). In addition to altering the sequence of the expressed pilin this process can also generate pilin genes that cannot be assembled or secreted (Haas & Meyer, 1986; Swanson *et al.*, 1986; Haas *et al.*, 1987; Manning *et al.*, 1991). Non-functional *pilE* genes can also be generated by deletions between repeated sequences within the locus (Hill *et al.*, 1990) and by duplication of the 3′ portion of the gene (Haas *et al.*, 1987). Both processes generate cells that are reversibly, through further recombination or deletions, piliated or unable to produce functional pili. The sequence alterations are associated with alterations in the substitutions of the pili with sugars (Virji *et al.*, 1993b) and therefore exchange of cassettes phase varies this modification by alteration of the substrate. Substitution with the

phosphorylcholine epitope is also independently phase varied by a mechanism that does not require pilin variation (Weiser *et al.*, 1998b). In addition, each of the *pilC* genes is independently phase variable. When neither *pilC1* nor *pilC2* are expressed the cells are unpiliated. When only *pilC2* is expressed the cells are piliated but not adherent. When *pilC1* is expressed (with or without *pilC2*) cells are piliated and also adherent. However, there are additional complications. In *N. gonorrhoeae* strains in which both *pilC1* and *pilC2* have been knocked out, cell lines that are initially unpiliated can revert to piliated (non-adherent) phenotypes in association with the expression of a novel 70 kDa protein, the role of which in non-mutants is unknown (Rudel *et al.*, 1995b). In addition, in strains in which *pilT* has been knocked out (a gene required for pilus mediated twitching motility) PilC is no longer required for pilus biogenesis (Wolfgang *et al.*, 1998). Finally, although not highlighted in the literature, there is important sequence difference between the pilus sequences of *N. gonorrhoeae* strains MS11 and P9 (Meyer *et al.*, 1984; Perry *et al.*, 1987) such that in the former there is a homopolymeric tract of 8 cytosines that is absent in the latter. This suggests that some forms of pili possess an additional mechanism of phase variation mediated by variation in this homopolymeric tract. Through the combination of these processes neisserial pili are highly variable and pleomorphic structures and the combined effects of PilC mediated phase-variation and alterations in the *pilE* gene have been implicated in transcellular passage of *N. gonorrhoeae* (Ilver *et al.*, 1998).

1.3.6.3 Phase variation of lipopolysaccharide structures

Like *Haemophilus* spp., *Neisseria* spp. have LPS that lacks O-side chain extensions and displays phase-variation. These species' LPS also share antigenic similarities (Virji *et al.*, 1990) which may reflect similar functional roles in these organisms which share a similar environmental niche and some features of associated infections. Different LPS phenotypes confer different properties upon the bacteria in addition to acting as a potential source of antigenic variation in immune evasion. Initially studied in *N. gonorrhoeae*, LPS was found

to be size and antigenically heterogeneous within a single strain (Apicella *et al.*, 1987; Schneider *et al.*, 1988; Weel *et al.*, 1989) and LPS phase variation has been observed during experimental infection of human volunteers (Schneider *et al.*, 1991). An association between LPS phenotype and invasive disease was found through the investigation of carriage and epidemiologically related invasive isolates collected from three countries (Jones *et al.*, 1992). It was found that LPS immunotype L3,7,9 was present in 97% of case/invasive isolates (13% also expressing L1,8,10) and 70% of colonising strains expressed L1,8,10 (20% also expressed L3,7,9). In addition an association between capsule expression and LPS immunotype was identified. In carrier isolates the loss of the capsule was associated with L3,7,9, whereas the capsulate isolates tended to express L1,8,10 epitopes more frequently than the non-groupable/acapsulate isolates (74% vs. 30%). In this study it is not clear how many sub-cultures these isolates had undergone prior to testing but it is likely that these differences were even greater in the initial isolates (see section 3 for illustrations of the influence of repeated culture on the population composition of phase variable characteristics). This study was followed by a study of bacteria with differing LPS immunotypes in an infant mouse intranasal infection model (Mackinnon *et al.*, 1993). This study confirmed L3,7,9 as a determinant of invasive disease consistent with the previous study. However the appropriateness of the model must be questioned because the predominant colonising strain was also L3,7,9 which differs from the human isolate study results. Meningococcal disease is a very rare event in colonised people, so any model in which the normal outcome is invasive disease is a poor representation of the steps that determine the progression from colonisation to infection. Indeed, one interpretation of this study is that the mouse model does not reflect human infection in this context. Organisms expressing the sialylatable phenotype (see below) L3 require pili to mediate adhesion to host cells whereas Opc (see below) mediated invasion only occurs in the context of more truncated (L8) LPS.

The LPS of *N. gonorrhoeae* can be externally modified by the sialylation of the terminal sugar residues (reviewed in Smith, 1991), an event that occurs *in vivo* (Apicella *et al.*, 1990; Parsons *et al.*, 1990) and which confers resistance to killing and opsonisation by normal human serum (Parsons *et al.*, 1989; Gill *et al.*, 1996). It also impedes adherence, and hence uptake and killing by neutrophils (Kim *et al.*, 1992; Rest & Frangipane, 1992). The mechanism by which serum sensitive and resistant phenotypes varied was found to be due to phase variation of the LPS side chain that was substituted with the sialic acid (van Putten, 1993) which was subsequently shown to be due to variation of expression of *lsi-2* (also called *lgtA*) (Danaher *et al.*, 1995). This study showed that the phase variation affected the behaviour of the cells in cell culture. LPS variants with little sialic acid invaded efficiently but were susceptible to complement mediated killing. Phase variation resulted in highly sialylated, equally adherent but entry deficient bacteria, which were resistant to killing by antibodies and complement due to altered complement activation and also masking of some LPS and protein epitopes (Poolman *et al.*, 1988; Judd & Shafer, 1989; van Putten, 1993; de la Paz *et al.*, 1995). The situation in *N. meningitidis* is similar but not identical.

Following inoculation of human volunteers with *N. gonorrhoeae* there is an 'eclipse period' during which bacteria cannot be cultured that suggests a 'programmed' intracellular phase during infection. During the early stage of infection only non-sialylatable LPS variants were isolated, whereas after the development of an inflammatory response the sialylatable phenotypes are recovered (Schneider *et* al., 1991). A similar obligate intracellular stage has not been demonstrated in *N. meningitidis*, hence the significance of interconversion between invasive and non-invasive phenotypes may differ in the two species. *N. meningitidis* also differs in that it can synthesise the substrate required for sialylation and hence does not need to acquire it from the environment. It is however similar to *N. gonorrhoeae* in the way that variation in sialylation is determined by phase variation of the lacto-N-neotetraose component of the LPS. Expression of this

epitope affects interactions with host cells including invasion of epithelial and endothelial cells (Virji *et al.*, 1995) and neutrophils (McNeil & Virji, 1997), and also neutrophil activation (Klein *et al.*, 1996). When the capsule and the longer, sialylatable LPS are not expressed then other proteins, in addition to pili, affect neisserial interactions with host cells (Virji *et al.*, 1992a & 1993a).

1.3.6.4 Phase variation of Class 5 surface proteins (Opa and Opc)

Neisseria spp. possess a family of surface proteins called Opa proteins (formerly called PII in *N. gonorrhoeae*) because they affect the opacity of the colonies during culture (Swanson, 1978). The *opa* genes in *N. gonorrhoeae* and *N. meningitidis* are phenotypically and structurally similar (Kawula *et al.*, 1988). Early reports showed that these proteins reduce sensitivity to serum and to some antibiotics, increased adhesion to buccal epithelial cells and that different Opa proteins interacted to different extents with leukocytes (Lamden *et al.*, 1979). The class 5 proteins (Opa and Opc) are particularly immunogenic and are presumably important targets for immune responses to *N. meningitidis* (Wiertz *et al.*, 1996). The Opa proteins are divergent, particularly in two C-terminal hypervariable regions, and are independently phase-varied (Sparling *et al.*, 1986; Stern *et al.*, 1984 & 1986). *N. meningitidis* typically possess 3 or 4 such genes (Aho, 1989; Aho *et al.*, 1991) whilst strains of *N. gonorrhoeae* have been reported to possess as many as 11 *opa* genes (Connel *et al.*, 1990; Bhat *et al.*, 1991; Dempsey *et al.*, 1991) and unlike the multiple partial copies of the pilin genes each one can be expressed. One of these studies also demonstrated that *opa* genes are capable of non-reciprocal inter-*opa* recombination to generate novel *opa* variants in a fashion similar to the pilin cassettes (Bhat *et al.*, 1991). This type of variation alone or in combination with horizontal transfer has led to the presence of at least 7 Opa variants in different strains within what is considered to be a clonal population of serogroup A *N. meningitidis* (Achtman *et al.*, 1988). Variation in the expression of Opa proteins has been shown to occur *in vivo* in the nasopharynx of *N.*

meningitidis carriers (Woods & Cannon, 1990). Early studies suggested that Opa proteins played a role in mediating *N. gonorrhoeae* interactions with host cells (Sugasawara *et al.*, 1983; Besson & Gotschlich, 1986; Makino *et al.*, 1991) and the adhesion and invasion phenotypes with which they are associated are also present when they are expressed in *Esch. coli* (Belland *et al.*, 1992; Simon & Rest, 1992; Gorby *et al.*, 1994). Subsequently it has been shown that the interactions between *N. gonorrhoeae* and host cells differ between cells expressing different Opa proteins (Fischer & Rest, 1988; Weel *et al.*, 1991a). Variant *N. gonorrhoeae* expressing different Opa proteins were shown to differ in attachment and mucosal damage caused to fallopian tube tissues (Dekker *et al.*, 1990). Experiments in which single *opa* genes are knocked out demonstrate that some affect both adhesion and invasion (Kupsch *et al.*, 1993) whilst others influence only invasion (Waldbeser *et al.*, 1994), and variant expression can also affect the cell tropisms of strains (Kupsch *et al.*, 1993). It should be remembered that the other *opa* genes remained intact in these studies. In a study using variants expressing one of 3 Opa proteins (or Opc) differences in adhesion and invasion were demonstrated for different epithelial cell lines, endothelial cells (Virji *et al.*, 1993a) and monocytes (McNeil *et al.*, 1994). Study of the receptors to which Opa proteins bind have shown that different Opas can interact with heparin sulphate proteoglycan which can confer binding to non-polarised cell lines and uptake through a number of mechanisms and to various members of the CD66 / CEA family of receptors (reviewed in Naumann *et al.*, 1999). Taken together these experiments indicate that the variable expression of the Opa proteins, as with the variant pilins, would be expected to alter the properties of the bacteria with respect to their interactions with host cells in addition to providing a mechanism of immune evasion. Finally, their functions may interact with other phase varied surface structures – for example purified Opa proteins have been shown to interact with the varied components of LPS (Blake *et al.*, 1995).

Opc is only present in *N. meningitidis*, although an ORF with moderate homology has been described in *N. gonorrhoeae*, as has a previously unrecognised homologue within *N.*

meningitidis strains (Zhu *et al.*, 1998). The function of these homologues has yet to be demonstrated and it is premature to consider them to be orthologs. Opc is included in the Class 5 proteins of *N. meningitidis* (with the Opa proteins) with which it shares properties of phase variation and heat-modifiability indicating the multimeric nature of the assembled protein on the cell surface. Initially considered an Opa protein, Opc (formally OpaC) was found to differ from the other Class 5 proteins by N-terminal sequencing, amino acid composition, electrophoretic properties, immunogenicity in rabbits, immunological cross-reactivity, its conservation within a collection of related isolates and the nature of its phase variation (Achtman *et al.*, 1988). Sequencing revealed that Opc shares only 22% sequence identity with Opa protein (Olyhoek *et al.*, 1991). In a fashion similar to the Opa proteins, Opc increases adherence of unencapsulated *N. meningitidis* to Chang and human umbilical vein derived endothelial cells (HUVECs). It also contributes to cell tropisms, being more efficient at mediating the interaction with HUVECs and neutrophils than with the Chang epithelial cells and monocytes – the reverse of what is observed with some Opa proteins (Virji *et al.*, 1992a & 1993a; McNeil & Virji, 1997). In comparisons of Opa proteins and Opc, Opc has been shown to be the most important determinant of cellular adhesion and invasion and promotes adherence to epithelial cells as effectively as OpaB – the most adherent Opa protein (Virji *et al.*, 1992a & 1993a). Monoclonal antibodies directed against Opc but not Opa proteins inhibit these interactions. Opc is discussed in more detail in the introduction to Chapter 7.

1.3.6.5 Phase variation of other neisserial surface proteins

Other structures whose role has been less directly associated with interactions with the host and which might contribute to invasive disease are also phase-variable. The class 1 protein, PorA, is a pore forming protein with cationic selectivity (Tommassen *et al.*, 1990) that has been sequenced (Barlow *et al.*, 1989) and its variation between strains is the basis of serological subtyping (serotypes being based upon the class 2 or class 3 proteins encoded

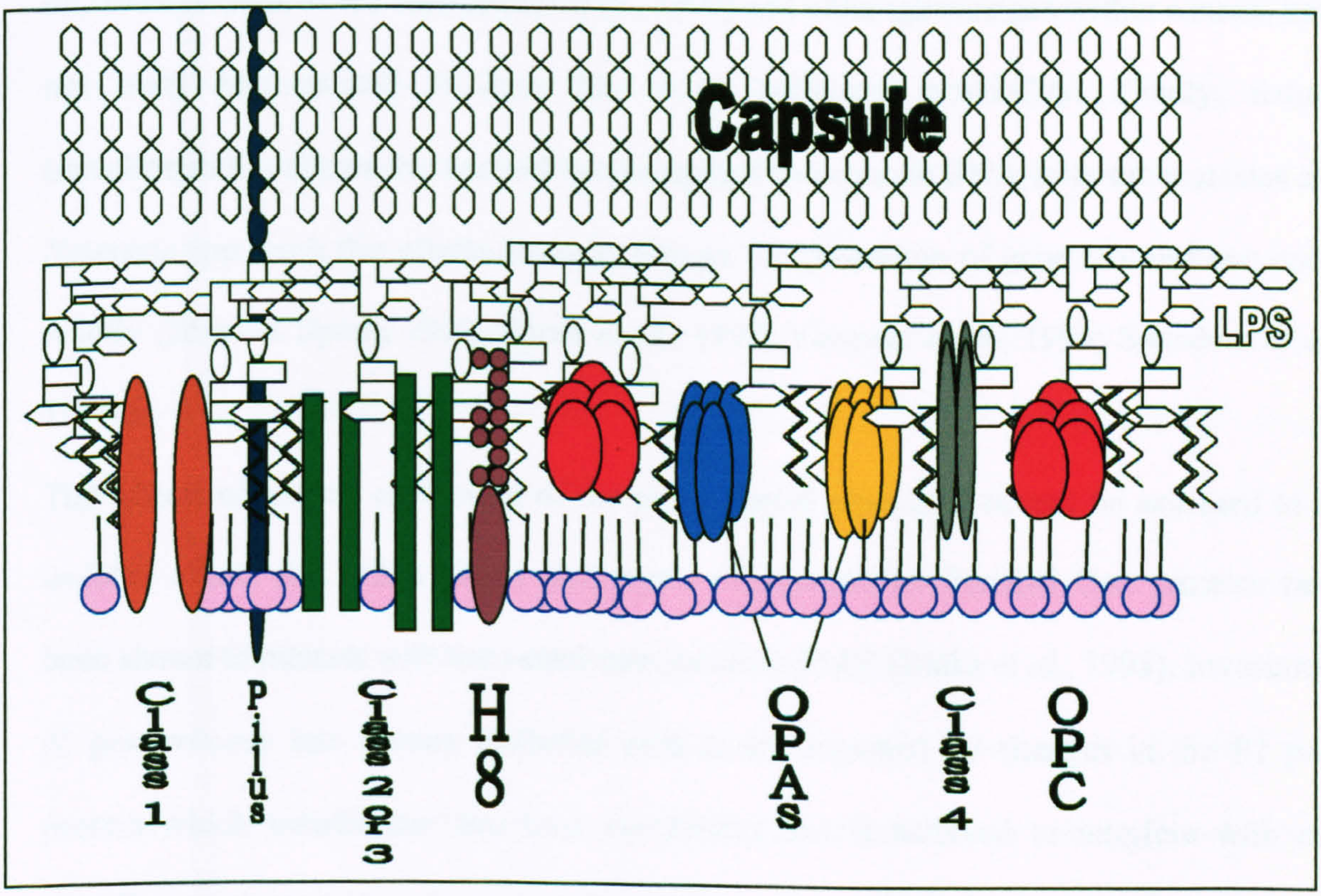
by *porB*). That antibodies formed to this protein place a selective pressure on organisms expressing PorA is suggested by studies that show subtype specific monoclonal antibodies directed against the subtype-specific epitopes are effective in bactericidal assays and are protective in animal models (Saukonen *et al.*, 1987 & 1989). The class 1 protein shows variation in expression in clinical isolates and is phase variable (Poolman *et al.*, 1980; Hopman *et al.*, 1994; van der Ende *et al.*, 1995).

While all bacteria require iron for growth free iron is present in limiting concentrations in the host (reviewed in Weinberg, 1978). Whether iron binding represents a 'virulence determinant' or not is a debatable issue since this requirement is common to virulent and avirulent bacteria. Haemoglobin utilisation has been found to be phase variable in *N. gonorrhoeae* through altered expression of the outer membrane protein HpuA (Chen *et al.*, 1996 & 1998). Variation in expression of these surface structures may simply contribute to immune evasion or they may have other, as yet uncharacterised, functions in cellular interactions. Finally there is evidence, not only that invasion of primary nasopharyngeal epithelial cells by *N. meningitidis* involves phase variation of multiple surface antigens and that Opa and Opc affect tissue tropisms, but that there are still other variable proteins, including one of 28-kDa, that are associated with the invasive potential of *N. meningitidis* (de Vries *et al.*, 1996).

1.3.6.6 The effect of phase variation on the composition of the neisserial cell surface

The surface of *N. meningitidis* is represented in a cartoon diagram in figure 1.1. The surface that *N. meningitidis* presents to its environment and to the hosts' immune responses is a highly fluid and dynamic structure. The components thought to affect host cell interactions, including capsule, pili, Opa proteins, Opc, LPS, porins and iron binding proteins, are all phase-variably expressed. As described above, further diversity within some of these structures is also considerable e.g. pili, Opa proteins and LPS. In this context it is noteworthy that one of the few relatively stable surface structures, Class 4 proteins,

Figure 1.1 Diagram representing the major surface structures of *Neisseria meningitidis*. Including the Class 1 to Class 5 proteins (Opa and Opc), capsule, LPS, and pilus.



ellicits blocking antibodies which may serve to protect the organism from the immune response (Munkley *et al.*, 1991). Even the structures that are not phase-variable cannot be considered to be stable at the level of the population. The H8 protein (or Lip) is surface exposed and found primarily in pathogenic members of *Neisseria* (Cannon *et al.*, 1984) and is believed to have a conserved sequence. Whilst it is not known to be variably expressed, the antigen is in some strains either absent or sequestered (Robinson *et al.*, 1987). It is composed of highly repetitive sequence motifs that varies in size and repeat component numbers between strains (Cannon *et al.*, 1984; Hitchcock *et al.*, 1985; Baehr *et al.*, 1989; Woods *et al.*, 1989; Pettit *et al.*, 1990) and although changes within a strain have not been demonstrated to date this is an additional possibility. Finally, natural transformability allows this species to take up and incorporate DNA from other strains and *Neisseria* spp. such that allelic diversity though the generation of gene mosaics can occur readily (Zhou & Spratt, 1992; Spratt *et al.*, 1992; Vazquez *et al.*, 1995; Saunders *et al.*, 1999a).

The effects of altered expression of the phase varied structures cannot be assumed to be isolated events that affect single phenotypic characteristics. Purified Opa proteins have been shown to interact with the varied components of LPS (Blake *et al.*, 1995). Invasion of *N. gonorrhoeae* into human epithelial cells is accompanied by changes in the P1 pore protein which translocates into host membranes and is believed to interfere with cell signalling pathways (Lynch *et al.*, 1984; Haines *et al.*, 1991; Weel & van Putten, 1991; Weel *et al.*, 1991a & b). One mechanism whereby sialylation of LPS might affect bacterial uptake but not adhesion is by interfering with PI mediated function. Consistent with this, LPS phenotype affects the immunoaccessability of PI (Poolman *et al.*, 1988; Judd & Shafer, 1989) which particularly affects protein epitopes that are not on the most surface exposed loops (de la Paz *et al.*, 1995). It is also possible that interactions between different neisserial proteins and structures affects cellular interactions. For example, one of the porins, PorB, in *N. gonorrhoeae* has been shown to influence Opa-dependent invasion of

cultured epithelial cells (Bauer *et al.*, 1999) and there are similarities between Opa proteins and the cellular receptors to which LPS binds on host cell surfaces (Porat *et al.*, 1995).

1.3.7 Phase variable restriction modification systems

The phase variable bacterial components described in the preceding sections are genes for which it is possible to discern possible roles in the bacterium:host interaction. In addition to these there are several restriction modification systems which are either phase variable or are associated with repeats that are likely to confer phase variability. A phase variable type I restriction-modification system has been described in *Mycoplasma bovis* that is associated with variable resistance to bacteriophage (Bhugra & Dybvig, 1992; Dybvig & Yu, 1994). There are two potentially phase variable restriction modification systems (one type I and one type III) described in *H. influenzae* (Hood *et al.*, 1996; van Belkum *et al.*, 1997a; also see section 9), one type III system in *N. gonorrhoeae* (Belland *et al.*, 1996), and two type I systems in *P. haemolytica* (Ryan & Lo, 1999). The role of phase variation of these genes in these species is unknown.

1.3.8 Common themes in the phase variable systems

The presentation of phase variation in sections 1.3.1 to 1.3.6 has presented a review of the literature on phase variation in a broadly species based and approximately historical order. Each species group has been presented so that the potential complexity conferred by multiple phase variable systems present in a single species group can be appreciated. The historical order also highlights the development of the field. This progresses from the phase variation associated with relapsing fevers and *Salmonella* typing, through to the use of whole genome sequences and the complex role of phase variation in species which make extensive use of repeats to vary many determinants of host interaction. Each species group has been used to highlight different aspects of this progression. However, there are common themes that are not emphasised by this approach. Phase variable structures

frequently play central roles in bacterial adhesion. The alteration of substrate specificity increases the range of microenvironmental niches that are available to an organism that can phase vary these determinants, an aspect of phase variation that can be easily related to the contingency gene model. The ability to switch off these adhesins might also play a role in release from host cell surfaces and subsequent dissemination. Structures mediating adhesion to host cells are surface exposed and often extended structures, and are therefore potential immunological targets. The ability to switch off these genes, as well as many other cell-surface located potential antigens, can also be seen to be potentially important in this context. Immune evasion in *Borrelia* spp. (section 1.3.1) and *H. somnus* (section 1.5.4) are particularly good examples. This argument can also be applied to determinants of motility which may be important in moving to establish colonisation in new locations but where continued expression may be detrimental.

The expression of surface determinants which may not be involved in adhesion but which are involved in immune evasion such as the bacterial capsules, sialylation of LPS and the expression of phosphorylcholine, are also phase varied. The expression of these determinants may be important in bacterial persistence and dissemination. However, their expression under other circumstances may impair the action of bacterial surface components important for host cell interactions. In this instance there will be different situations in which ON and OFF phenotypes are adaptive.

It need not be assumed that each alternate surface structure confers a unique niche adaptation. It is likely that there is also substantial functional redundancy in some instances. For example in the case of the variable protein families such as the multiple variable surface proteins of the *Mycoplasma* spp. (section 1.3.4), the iron binding proteins of *Haemophilus*, or the *Opa* proteins of *Neisseria*, there may be substantial functional redundancy. This may also be the case for some of the many potential LPS phenotypes of *Haemophilus* and *Neisseria*.

The majority of phase variable genes recognised to date can be divided into a small number of functional categories: surface proteins, enzymes modifying surface proteins, proteins involved in the biosynthesis of non-protein surface structures, toxins and restriction enzymes. With the exception of the restriction enzymes (section 1.3.7) the roles of these types of protein are established in bacterial interactions with their hosts and in virulence. There is a strong association of phase variable genes with functions important in adaptation to variable aspects of the environment.

1.4 Phase variation as a virulence determinant in *Neisseria meningitidis*

There are several aspects of *N. meningitidis* that make study of the processes involved in pathogenesis and the identification of vaccine candidates difficult. One central issue is that it is hard to delineate the functions of bacterial components that are involved in benign interactions and to differentiate them from those that contribute to the infection process. *N. meningitidis* is naturally transformable which enables the population as a whole to readily exchange genetic information to generate novel solutions to new selective pressures and to use multiple copies of genes within clonal populations to generate new variants. Finally, *N. meningitidis* have an enormous capacity to stochastically switch on and off contingency genes to express them in varied combinations within a clonal population over time.

N. meningitidis is most frequently associated with humans as a harmless commensal organism colonising the nasopharynx. The proportion of the population colonised varies over time and outbreaks can be associated with increases in the proportion of the population that are colonised. The factors that affect levels of colonisation include climatic conditions, variations in the strains present, and host factors such as living conditions. Reflecting this the incidence of meningococcal meningitis shows seasonal variation, can show epidemics following alterations in bacterial surface components, and rises when groups of susceptible individuals are placed into confined conditions which favour transmission (e.g. army recruits and students). However, even under these conditions

invasive disease is a relatively infrequent event in comparison to the prevalence of benign colonisation and is not invariably seen under the conditions associated with outbreaks.

Some of the most difficult and important questions with respect to infections caused by bacteria that are part of the 'normal flora' remain to be answered, particularly: who, when, and why (the 3Ws) will an individual suffer an infection caused by a particular species or strain.

That the identification of what have traditionally been considered to be virulence determinants in these organisms does not, in itself, help us to answer these questions is illustrated by some observations. First, a full complement of virulence genes is present in those bacteria that do not cause infection and from which the disease causing organism is derived. Second, genes that are considered to be virulence genes are not universally present in those strains that do cause disease. Finally, well established virulence determinants are present in non-pathogenic strains and species.

To illustrate these points with examples relevant to *N. meningitidis*:

1. In an outbreak of meningitis there is an increase in the proportion of individuals in a community who are colonised with the strain that is responsible for the outbreak. The strain that causes the infection is (to date) indistinguishable from those colonising the other individuals in the population.
2. Opc has been studied intensively both as a virulence determinant and as a vaccine candidate (see sections 1.3.4 and 7.1) and can reasonably be considered to be important in pathogenesis. However, it is not present in ET-37 complex serogroup A meningococci (nor is it present in around 30% of groups B strains (see section 7.3)) and strains that lack *opc* are fully virulent.
3. Opa proteins are present in the non-pathogenic species *Neisseria lactamica* where they are associated with pentameric repeats in the same way as they are in *N. meningitidis* and *N. gonorrhoeae* and are presumably also phase-variable as a result. Thus, neither the possession of Opa proteins nor the capacity to phase-vary their expression is

sufficient to confer virulence on neisserial spp., although it is possible that the binding characteristics of the Opa proteins in *N. meningitidis* and *N. gonorrhoeae* might differ from those in *N. lactamica*. However, the ability of the neisserial species to exchange DNA of homologous genes suggests that such compartmentalisation of specific Opa genes to *N. meningitidis*, when its ecological niche and the surface with which it interacts is so similar to that inhabited by *N. lactamica*, seems unlikely.

One possibility is that the principal determinants of the 3Ws are host factors and that each individual is infected with an organism of equal propensity to cause infection. In clinical practice one of the most powerful questions in the diagnostic approach to an infectious diseases case is: 'What is the nature of this patient's immunocompromise?' This is not only useful when seeing patients with gross abnormalities such as those with haematological disorders and iatrogenically induced states related to therapy for transplants or autoimmune conditions. It is also useful when seeing patients with metabolic disorders, foreign bodies, those with disease or degeneration induced immunologically privileged sites, those who have undergone surgery or anaesthesia, and those who are immunologically naive such as children or travellers. Combined with a detailed knowledge of the organisms that are present in the potential sources of infection this approach can be helpful in almost all circumstances.

However, this argument cannot be used in reverse. Not every patient who is bitten by a dog or with liver cirrhosis becomes infected with *Capnocytophaga*. Not every patient with diabetes or being treated with aggressive chemotherapy develops superficial or deep infections with *Staph. aureus* although all are colonised and/or exposed to this organism. Wound infections are common – but they are not universal. There must be other aspects of the bacterium : host interaction which determine the 3Ws.

In the context of meningococcal meningitis this perspective raises two central questions: 1. Why are host factors insufficient to identify the 3Ws? and, 2. How can colonisation with

what is ostensibly the same bacterium lead to such different outcomes in different individuals?

With respect to the first question:

Patients who have terminal complement deficiencies are at increased risk of developing invasive meningococcal disease. However, this does not occur in all such patients, nor to all splenectomised individuals who are similarly susceptible, even though the possibility of them never being colonised by disease associated strains is unlikely. Infection often occurs in adult life or years after splenectomy and this is unlikely to be due to the first potentially virulent *N. meningitidis* strain to colonise the patient. This argument can also be applied to young adult cases of meningitis. These individuals will probably have previously encountered *N. meningitidis* several times and have not succumbed – so the development of infection would appear to be strain or clone specific, or determined by additional co-factors which increase host susceptibility. Further, in outbreaks of meningitis, although the circulating clone colonises a large number of individuals – 'close contacts' have a higher risk of developing infection than more 'remote' individuals. This is not restricted to related contacts and therefore cannot be explained entirely on the basis of shared characteristics within families.

With respect to the second question:

Phase variation is a property of these bacteria that is capable of mediating clonal variation that could affect virulence within a strain that would be indistinguishable using existing methods. Interestingly it is shared by other bacterial species, such as *H. influenzae*, that normally colonise their hosts without causing infection but which do so rarely and apparently at random in a way reminiscent of *N. meningitidis*. Remarkably, as described in the previous section, this process affects the expression of almost all of the surface structures of *N. meningitidis* that have been suggested to be important in virulence. Their variation and the novel combinations of determinants that it stochastically produces may have the capacity to generate exactly the pattern of meningococcal disease that is seen in

the population. It can be argued that it is not the possession of the genetic information for any particular bacterial structural component or product that confers virulence on this type of infectious agent. Instead, it is the way that their expression is controlled and the timing and context of that expression that determines the progress of infection in the potentially susceptible host. It also provides a means by which a single clonal bacterial population can generate colonising populations that differ from host to host. Indeed, the capacity to use phase-variation confers great flexibility on a population and the fact that it might also represent the basis of the organisms capacity to cause infection may be entirely incidental. In this model the invasion can be seen as a mistake. The organism has the capacity to phase-vary its surface structures. It probably, through niche fitting and selection, normally adopts a limited repertoire of gene combinations that fit it to different microenvironments, such as those present in the nasopharynx, that represent the typical starting conditions for diversification. The fact that this process has the capacity to generate more virulent clones through the expression of structures in particular combinations or in sites in which they are not normally expressed is unfortunate. As long as the adaptive advantage of this flexibility provides a selectable advantage in the normal transmission – colonisation cycle, then the loss of a small percentage of colonised hosts will not provide any significant selective pressure against this process and infection.

1.5 Mechanisms mediating phase-variation

The reversible switching mechanisms that mediate phase-variation fall into a number of categories:

1. non-reciprocal exchange (between silent and expressed loci)
2. sequence inversions (of promoter regions and reading frames)
3. insertion of mobile elements
4. insertions or deletions within repeats located within coding regions (that alter the reading frame of the protein)

5. insertions or deletions within repeats located within promoters (that alter transcription)
6. DNA modification

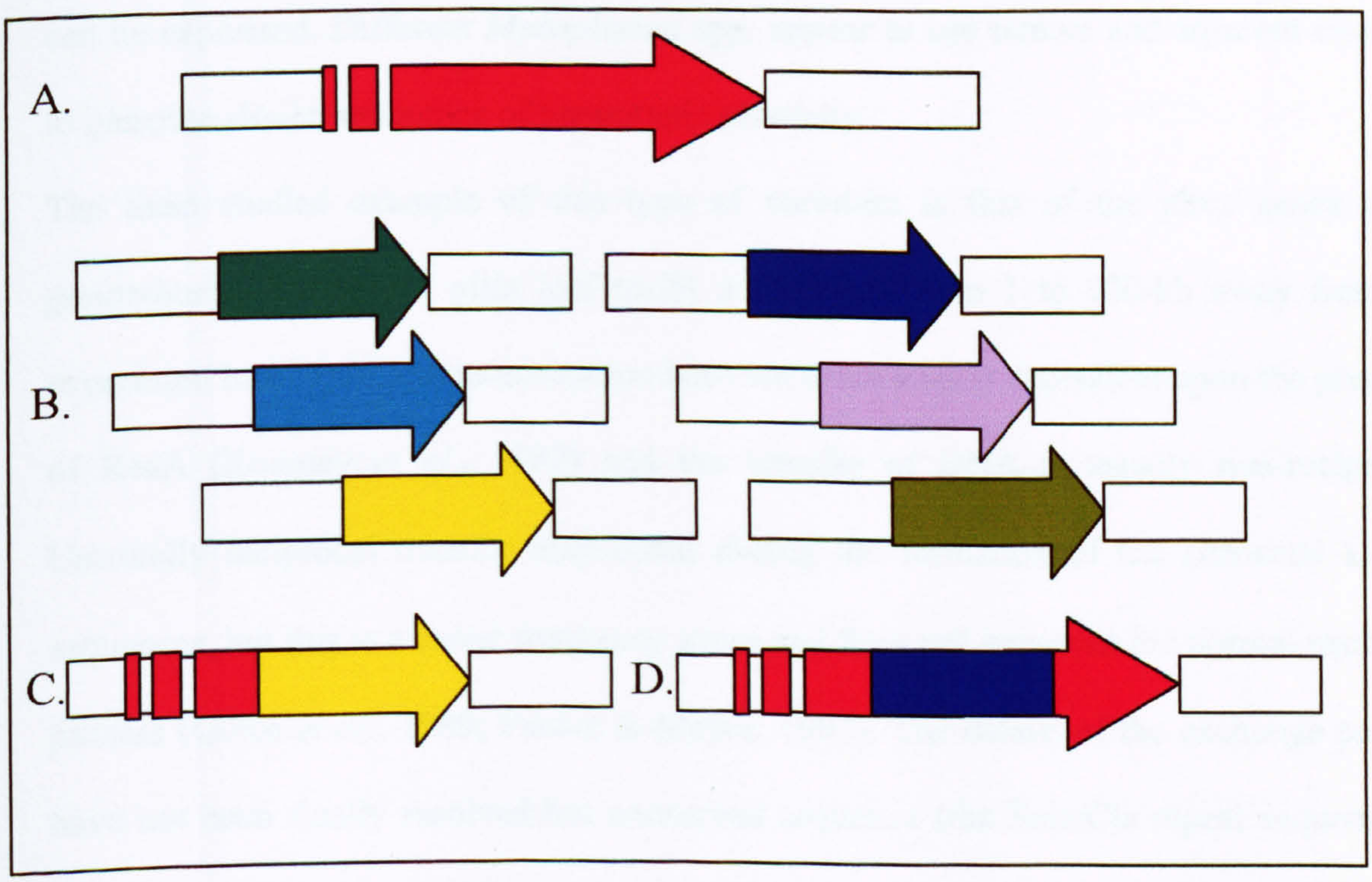
1.5.1 Non-reciprocal exchange

This mechanism of phase variation involves the movement of reading frames between sites within the genome where they are or are not associated with a functional promoter sequence (figure 1.2).

The variations of the VMPs in *B. hermsii* (section 1.3.1) correlate with DNA rearrangements which involve duplication of the expressed sequence at a site remote from the first (silent) copy (Meier *et al.*, 1985). This is due to the movement of 'silent' copies of the antigenic variants to a single expression locus. The silent copies of the genes and the expression locus are present upon separate multicopy linear plasmids (Plasterk *et al.*, 1985). *B. burgdorferi* uses a similar mechanism of phase variation but the genomic organisation differs. In *B. burgdorferi* the silent cassettes are located together in a single locus that is located approximately 200-bp 5′ of the expression locus. Each silent gene and the expression locus are delimited by a 17-bp direct repeat. (Zhang *et al.*, 1997; Kawabata *et al.*, 1998). Variation in the expression locus has been shown to be through a unidirectional process such that the sequence and organisation of the silent loci are not altered (Zhang & Norris, 1998). It is unclear at this time whether this process truly generates reversible phenotypic changes or whether variation in the cross-over sites results in the continuous generation of novel expressed genes in *B. burgdorferi*. In *B. hermsii* the process is reversible. Interestingly, and consistent with this concept of 'contingency genes', particular phenotypes have been shown to be expressed in different hosts (Schwan & Hinnebusch, 1998).

Mycoplasma genitalium expresses an adhesion protein (MgPa) that is immunogenic and required for attachment of the cell to the host epithelium. Use of probes prepared from the gene encoding MgPa showed that certain regions of the gene were present in multiple

Figure 1.2 Diagram representing gene variation due to the the presence of multiple allelic copies of a gene. Only a single copy of the gene (A - represented in red) has a functional promoter sequence (represented by the hatched section). Recombination between the cassettes that do not have a promoter (B) and the allele with the functional promoter leads to variation in the expressed version of the gene. This can result in the exchange of the whole of the potential coding region from the point of recombination (C) or incorporation of a section of the recombined silent cassette (D).



copies at remote sites in the genome whilst others were present only once (Dallo *et al.*, 1991). Analysis of these other copies suggest that they do not encode expressible reading frames and a variety of mosaics have been identified in unrelated strains that suggest that the silent sequences provide a reservoir for exchange with the expressed locus (Peterson *et al.*, 1995). The extent to which these recombinatorial processes are reversible has not been determined. It is unclear as to whether generation of mosaics or the programmed transfer of entire cassettes is the usual mechanism of change involving these sites. Each of the examples described below in the section on sequence inversions can also be considered to represent the interconversion of a silent and expression locus. The example of the *vsa* genes (section 1.3.4) also includes a large element of re-organisation that alters the loci that can be expressed. Different *Mycoplasma* spp. appear to use remote and adjacent cassettes to generate similar outcomes of phenotypic plasticity.

The most studied example of this type of variation is that of the pilus genes of *N. gonorrhoeae*. The silent pilin loci (*pilS*) are located from 1 to 900-kb away from the expression locus (*pilE*). Recombination between these sites is dependent upon the presence of RecA (Koomey *et al.*, 1987) and the transfer of DNA is usually non-reciprocal. Unusually reciprocal transfer may occur during the formation of the abnormal L-pilin sequences, but this is a lower frequency event and does not represent the normal exchange process (Gibbs *et al.*, 1989; Facius & Meyer, 1993). The details of the exchange process have not been finally resolved but conserved sequence (the Sma/Cla repeat sequence) is needed for efficient recombination to occur from most *pilS* loci (Wainwright *et al.*, 1994). DNA exchanged by natural transformation can lead to events that resemble those seen during pilin variation (Norlander *et al.*, 1979; Seifert *et al.*, 1988; Gibbs *et al.*, 1989) but similar changes can be observed when transformation is prevented (Swanson *et al.*, 1990; Zhang *et al.*, 1992; Facius & Meyer, 1993). The rates at which the events occur *in vitro* favour non-transformation mediated processes as the most probable mechanism (Zhang *et al.*, 1992). There is little functional distinction between these two alternatives since most

donor organisms will be from the same clonal lineage as the recipient – although the processes that might increase or decrease the frequency with which the events occur are potentially different. The extent to which any particular version of pilin can be regenerated, and thus fulfil the definition of phase variation, is unclear. It is likely that the original model: that each expressed gene is a mosaic of silent sequences from different *pilS* genes, is correct (Haas & Meyer, 1986). The contribution of these processes to phase variation is through the generation and then elimination of functionally abnormal pilin variants, such as due to deletion or amplification of sections of the *pilE* locus that are subsequently corrected by recombination (Hill *et al.*, 1990; Manning *et al.*, 1991) – such that the expression of pili is phase varied even if the sequence of the expressed pilin is altered.

1.5.2 Sequence inversions

The first example of a variable phenotype mediated by the inversion of a localised sequence region in prokaryotic genetics was the inversion of the G loop region of bacteriophage Mu and its correlation with the infectivity of the 'phage particle (Bukhari & Ambrosio, 1978; Kamp *et al.*, 1978). The bacterial phase variation events mediated by this type of mechanism are summarised in Table 1.1.

Table 1.1. Examples of phase variation mediated by sequence inversions.

Gene	Species	Reference(s)
H2 (SP)	*Salmonella* spp.	Zieg *et al.*, 1977 & 1978; Silverman & Simon 1980
type I fimbriae gene (SP)	*Esch. coli*	Eisenstein, 1981; Abraham *et al.*, 1985
pilin genes (SP)	*Moraxella bovis*	Marrs *et al.*, 1988
pilin genes (SP)	*Moraxella lacunata*	Marrs *et al.*, 1990
hsd1 (RM)	*Myc. pulmonis*	Dybvig & Yu, 1994
vsa (SP)	*Myc. pulmonis*	Bhugra *et al.*, 1995
mrpA (SP)	*Pr. mirabilis*	Zhao *et al.*, 1997

SP = surface protein, RM = restriction modification system protein

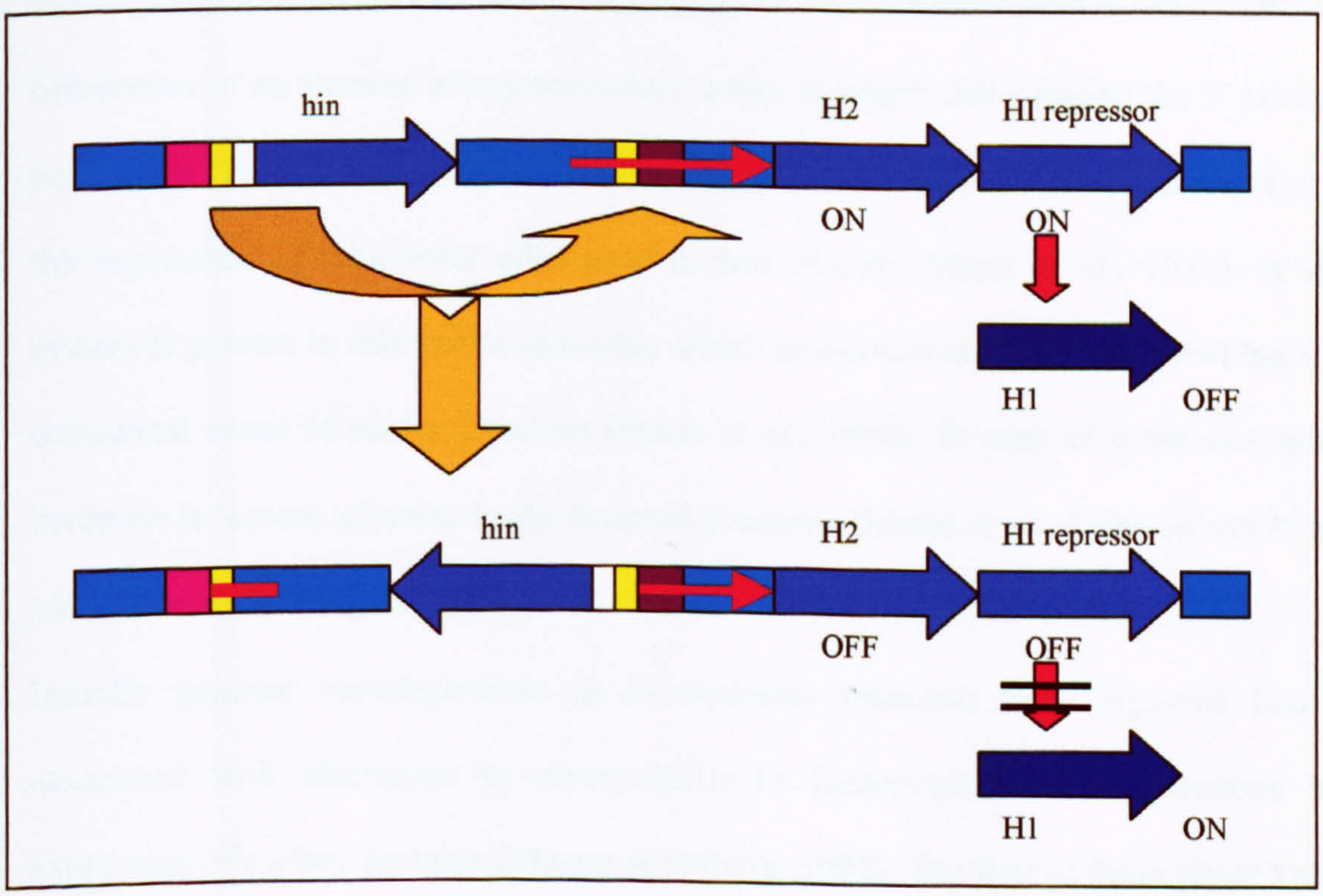
The classic example of bacterial phase variation mediated by sequence inversion is that of the promoter controlling the alternate expression of the H1 and H2 phase flagellar antigens

(Figure 1.3). The two flagellar genes map to different sites in the *Salmonella* genome (Lederberg & Edwards, 1953) and the expression of H1 and H2 are under the control of an element linked to H2 (Lederberg & Iino, 1956). This second element is a repressor, *rh1*, that prevents the expression of H1 and is co-ordinately expressed with H2 (Fujita *et al.*, 1973; Silverman *et al.*, 1979). Expression of H2 is controlled by the inversion of a 900-bp region adjacent to the H2 gene (Zieg *et al.*, 1977) (Figure 1.3). This includes the promoter and the sequence inverted in a site-specific and RecA independent fashion (Zieg *et al.*, 1978; Silverman *et al.*, 1979), the inversion being mediated by an invertase (*hin*) that is encoded within the inverted segment (Silverman & Simon, 1980; Zieg & Simon, 1980).

The type-1 fimbriae of *Esch. coli* are also under transcriptional control (Eisenstein, 1981) and are phase varied by a mechanism of promoter inversion (Abraham *et al.*, 1985) similar to that seen in the control of flagellar expression in *Salmonella*. It differs from the H2 *Salmonella* promoter described above in that the inverted element is smaller (314-bp), has different recombinase specificity (TTGGGGCCA), and the inverted sequence does not encode its own recombinase. The inversion of the promoter element is catalysed independently by two recombinases encoded by *fimB* and *fimE* and these differ in their biochemical specificity so that FimB can mediate inversion in both directions but FimE can only re-orientate the promoter from ON to OFF (Gally *et al.*, 1996; Kulasekara & Blomfield, 1999). In addition to this paired system of recombinases that biases the rate of switching in the direction from ON to OFF the rate and direction of switching is also environmentally responsive (Gally *et al.*, 1993). This is mediated by the binding of several accessory proteins including: integration host factor (IHF) (Dorman & Higgins, 1987; Eisenstein *et al.*, 1987), the leucine-responsive regulator (Blomfield *et al.*, 1993; Gally *et al.*, 1994) and the histone-like protein H1 (Kawula & Orndorff, 1991).

A similar system to that controlling type 1 fimbrial phase variation in *Esch. coli* is present in *Proteus mirabilis* (Zhao *et al.*, 1997). In this case expression of the mannose-resistant / Proteus-like (MR/P) pili, that are required for the development of acute pyelonephritis and

Figure 1.3 Diagram representing the phase variation of H1 and H2 flagellae in *Salmonella* mediated by inversion of the H2 promoter. The region shown with the curved orange arrow is inverted and the promoter sequence for H2 (indicated by the long red arrow) is disrupted leading to switching off of the expression of H2 and the H1 repressor. Loss of the H1 repressor leads to expression of H1. The H2 phenotype is regained by inversion of the inverted element to the original configuration.



are preferentially expressed in infection associated isolates in a murine model of infection, is controlled by the inversion of a 252-bp element that includes the promoter for the adjacent fimbrial gene. The gene for the invertase, *mrpI*, is located adjacent to the inverted element on the opposite side and orientation to the fimbrial gene. The expression of *mrpI* is not affected by the orientation of the element. Similar to the type-1 fimbrial switch, the process of sequence inversion is influenced by environmental conditions although the proteins and mechanisms are yet to be determined.

In each of these examples an element containing the promoter is inverted leading to altered expression. In *Moraxella bovis*, a bovine pathogen that causes eye infections, this situation is reversed. Single strains of *M. bovis* are capable of producing two types of pilin protein, called α and ß (Marrs *et al.*, 1985). Expression of each alternate pilin is determined by the orientation of an element of approximately 2-kbp in length that contains the 3′ portions of both genes. There is a single promoter that is adjacent to this inverted element and controls the expression of whichever pilin gene is downstream (Marrs *et al.*, 1988). A similar system is present in *Moraxella lacunata*, which is associated with humans where it is an occasional cause of ocular infection (Marrs *et al.*, 1990). In each of these examples the invertase is located adjacent to the inverted promoter (Marrs *et al.*, 1990; Rozsa & Marrs, 1991; Lenich & Glasgow, 1994).

Initially genome rearrangements in *Mycoplasma pulmonis* were reported that were associated with alterations in susceptibility to bacteriophage and alterations in the expression of surface proteins (Bhugra & Dybvig, 1992). The first of these phase variation phenomena to be defined in detail was associated with an inversion in the *hsd1* locus which encodes all three elements of a type I restriction-modification system (Dybvig & Yu 1994). Inversion of a 6.8-kbp element is associated with switching of restriction, modification and a 1000 fold increase in resistance to the mycoplasma virus P1. The element includes the S subunit (which confers sequence specificity) followed by the R subunit (required for restriction) and then the M subunit (the methylase required for both

restriction and modification). There is a second, divergent, S subunit on the opposite strand 3′ of the M subunit. The two S subunits share their 5′ portions such that they represent a 450-bp inverted repeat flanking the locus that includes the sites for the inversion. This example is notable for four reasons. Firstly the size of the inverted element is significantly larger than that seen in other examples. Secondly, the order of the R and M genes is reversed when compared with similar systems in other bacterial species. Thirdly, the invertase responsible for the reorganisation is neither contained within the inverted element or in the adjacent sequence. Fourthly, a second genetic reorganisation at a remote site was detected by a change in RFLP patterns associated with the inversion in the *hsd1* promoter.

The second site which changes with the *hsd1* locus provides an elegant and complex example of phase-variation mediated by inversions in the *vsa* locus encoding the variable V-1 surface antigens. The *vsa* locus contains a single promoter and several variant coding regions for alternate V-1 proteins. A complex series of site-specific inversions of a centrally located promoter and the flanking sequences alter which sequences represent expressed and silent genes (Bhugra *et al.*, 1995).

Each coding region has a sequence with a 34bp core consensus located at its 5′ end, which is the site at which the inversions occur. This includes two ORFs that divergently flank the promoter. Inversion of the promoter determines which of the two flanking sequences is expressed. Inversions within the regions either side of the promoter determine which ORF will be in the corresponding promoter flanking location.

There are other examples of phase variation, that are not mechanistically defined in detail, for which the available evidence suggests that similar inversion phenomena are involved. Phase variation of VspA in *Myc. bovis* has been associated with reversible changes in RFLP patterns similar to those seen in *Myc. pulmonis* (Lysnyansky *et al.*, 1996). There is also bi-directional variation between two antigenically different flagellar structures with different molecular weight composite flagellins in *Campylobacter coli* and *jejuni* (Harris *et*

al., 1987). This is associated with a genomic rearrangement consistent with a variation mechanism mediated by an invertible element (Guerry *et al.*, 1988).

1.5.3 Insertion of mobile elements

The reports of these mobile element insertions in the literature suggest that this may be a mechanism mediating phase-variation. However some of the reports are not without their difficulties, and it is not yet certain that this type of process mediates high-frequency reversible switching *in-vivo* in pathogenic bacteria. The reported genes are listed in Table 1.2.

Table 1.2. Examples of phase variation predicted to be mediated by the insertion of mobile elements.

Gene	**Species**	**Reference**
Vi antigen gene	*C. freundii*	Ou *et al.*, 1988
eps (CAP)	*Ps. atlantica*	Bartlett *et al.*, 1988
siaA (LPS & CAP)	*N. meningitidis*	Hammerschmidt *et al.*, 1996a

CAP = capsule biosynthesis gene, LPS = gene involved in LPS biosynthesis

The first report describing this mechanism was in a study of Vi antigen variation using a plasmid into which the variable gene (*viaB)* from *C. freundii* had been cloned (Ou *et al.*, 1988). This was then cloned into an *Esch. coli* strain HB101 background. What were described as 'reversible switching rates' in this paper were in-fact frequencies with which transformants expressed the alternate phenotype. Switching from Vi-positive to Vi-negative was observed in the absence of transformation, but never the reverse. Therefore this model does not display true phase-variation. The mechanism of inactivation of the *viaB* gene was insertion of an insertion sequence-1 like element into the cloned gene.

There is an example of phase variation that occurs in *Pseudomonas atlantica* that is associated with a genome rearrangement for which the most likely mechanism is movement of a mobile genetic element (Bartlett *et al.*, 1988). *P. atlantica* is a

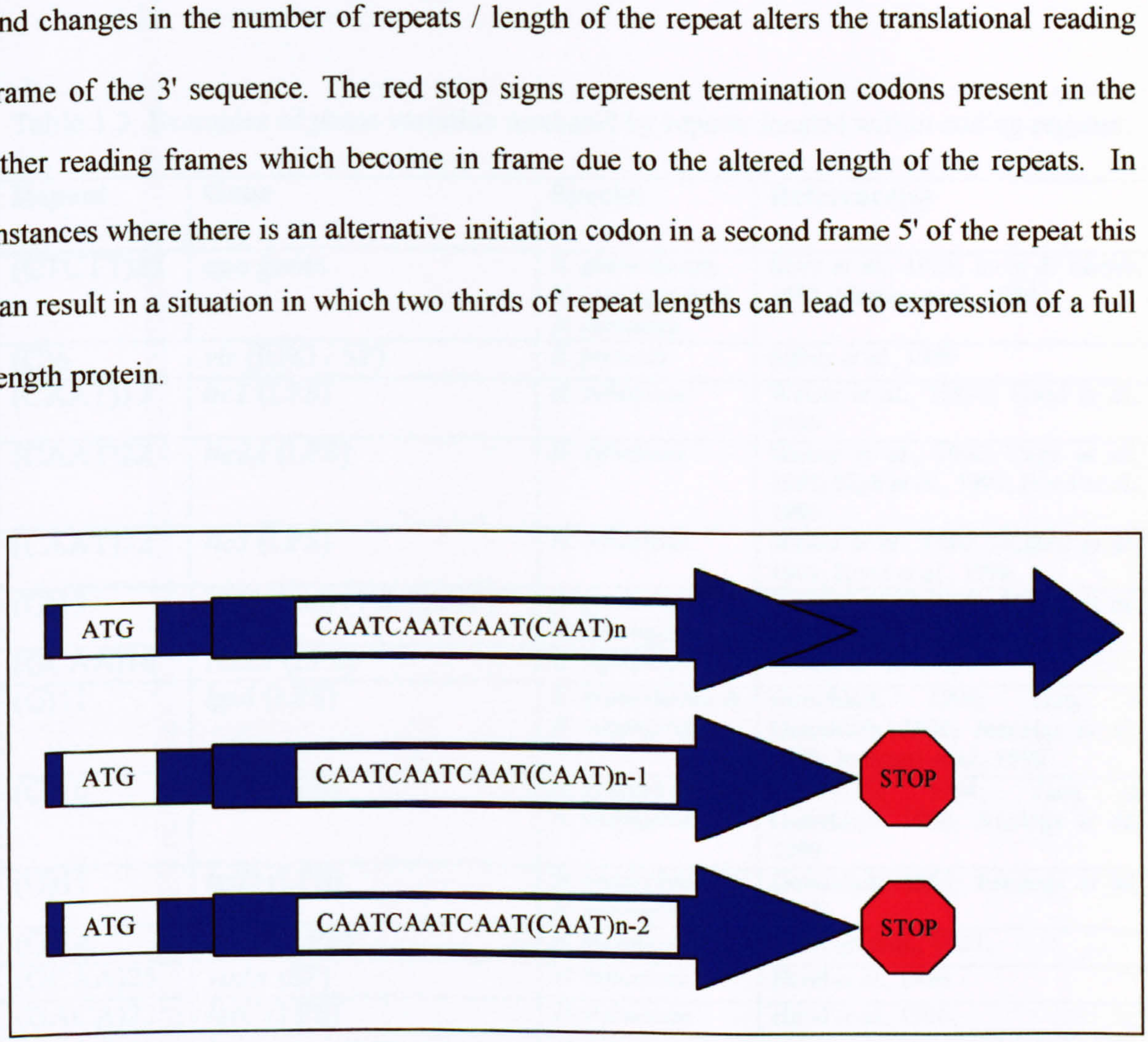
pseudomonad that is able to colonise a variety of marine environmental surfaces. One adhesive determinant is an extracellular polysaccharide which is reversibly phase variable. Phase variation is associated with a reversible non-duplicative insertion of a 1.2-kbp sequence which is present in multiple copies in the genome and an increase in the number of copies of the element which is RecA independent. These observations are consistent with the movement and precise insertion and deletion of an insertion sequence in the locus that controls polysaccharide expression. The sequence of the mobile element has not been determined.

The second example in a human pathogen, is the inactivation of the capsule biosynthesis gene *siaA* of *N. meningitidis* by insertion sequence 1301 (Hammerschmidt *et al.*, 1996a). In this experiment 2 of 30 unencapsulated variants selected from epithelial cell invasion assays were found to have an insertion sequence of 844-bp at position 587 of the reading frame. Four out of 5 repeat hybridisation experiments using a probe to the junction of the insertion sequence and *siaA* showed 10 to 15% of capsule negative variants to have the insertion. Regain of capsule in the IS1301 variants was observed and associated with precise excision of the insertion sequence. The difficulty with this example is that no capsule negative variants could be detected when a uniform starting population expressing capsule was used. The epithelial cell invasion selection experiment used to detect the variants were performed with subcultures from a single stock and all of the IS1301 associated variants may have been the product of a 'founder effect' resulting from a single occurrence of a rare event. Other experiments have shown that the majority of the capsule variants are generated by a different and more frequent event (see section 1.3.4) (Hammerschmidt *et al.*, 1996b).

1.5.4 Repeats within coding regions

Alteration in the length of simple oligonucleotide repeats that are composed of motifs that are not a multiple of 3 bases in length will alter the translational reading frame of the ORF

Figure 1.4 Diagram (using the tetrameric repeats present in the *lic* loci of *Haemophilus influenzae* as an example) representing the effect of altered repeat element length upon translation of an open reading frame. The repeats are located after the ATG initiation codon and changes in the number of repeats / length of the repeat alters the translational reading frame of the 3' sequence. The red stop signs represent termination codons present in the other reading frames which become in frame due to the altered length of the repeats. In instances where there is an alternative initiation codon in a second frame 5' of the repeat this can result in a situation in which two thirds of repeat lengths can lead to expression of a full length protein.



within which they are located. These repeats are frequently closely followed by termination codons in the two reading frames that do not generate a full length, functional protein. Repeat instability thus generates ON – OFF switching through the generation of abnormal mRNA during transcription. The affect of repeat length variation is illustrated in Figure 1.4.

This is the most frequently described mechanism of phase variation and genes that have been associated with this type of repeats are summarised in Table 1.3.

Table 1.3. Examples of phase variation mediated by repeats located within coding regions.

Repeat	**Gene**	**Species**	**Reference(s)**
(CTCTT)25	*opa* genes	*N. gonorrhoeae, N. meningitidis & N. lactamica*	Stern *et al.*, 1986; Stern & Meyer, 1987; Murphy *et al.*, 1989
(C)6	*vir* (REG / SP)	*B. pertussis*	Stibitz *et al.*, 1989
(CAAT)17	*lic1* (LPS)	*H. influenzae*	Weiser *et al.*, 1989b; Hood *et al.*, 1996
(CAAT)22	*lic2A* (LPS)	*H. influenzae*	Weiser *et al.*, 1990; Cope *et al.*, 1991; High *et al.*, 1993; Hood *et al.*, 1996
(CAAT)32	*lic3* (LPS)	*H. influenzae*	Weiser *et al.*, 1990; Maskell *et al.*, 1991; Hood *et al.*, 1996
(G)13	*pilC* genes (1 & 2) (SP)	*N. gonorrhoeae & N. meningitidis*	Jonsson *et al.*, 1991; Nassif *et al.*, 1994; Taha *et al.*, 1996
(GCAA)18	*lex2A* (LPS)	*H. influenzae*	Jarosik & Hansen, 1994
(G)11	*lgtA* (LPS)	*N. gonorrhoeae & N. meningitidis*	Gotschlich, 1994; Yang & Gotschlich, 1996; Jennings *et al.*, 1995; Jennings *et al.*, 1999
(G)10	*lgtC* (LPS)	*N. gonorrhoeae & N. meningitidis*	Gotschlich, 1994; Yang & Gotschlich, 1996; Jennings *et al.*, 1999
(G)11	*lgtD* (LPS)	*N. gonorrhoeae & N. meningitidis*	Gotschlich, 1994; Jennings *et al.*, 1999
(G)14	*lic2A* (LPS)	*N. meningitidis*	Jennings *et al.*, 1995
(GCAA)25	*yadA* (SP)	*H. influenzae*	Hood *et al.*, 1996
(GACA)22	*lgtC* (LPS)	*H. influenzae*	Hood *et al.*, 1996
(CAAC)36	Iron acquisition gene	*H. influenzae*	Hood *et al.*, 1996
(CAAC)20	Iron acquisition gene	*H. influenzae*	Hood *et al.*, 1996
(CAAC)18	Iron acquisition gene	*H. influenzae*	Hood *et al.*, 1996
(CAAC)20	Iron acquisition gene	*H. influenzae*	Hood *et al.*, 1996
(CAAC)15	FUN	*H. influenzae*	Hood *et al.*, 1996
(AGTC)32	Type III *mod* gene (RM)	*H. influenzae*	Hood *et al.*, 1996
(TTTA)6	FUN	*H. influenzae*	Hood *et al.*, 1996
(C)7	*siaD* (CAP)	*N. meningitidis*	Hammerschmidt *et al.*, 1996b
(NNNNN)n	*ngo*X.M (RM)	*N. gonorrhoeae*	Belland *et al.*, 1996
(CAAT)32	*lex2B* (LPS)	*H. somnus*	Inzana *et al.*, 1997
(CACAG)58	*alxA* & Type I restriction	*P. haemolytica*	Highlander & Garza, 1996;

	enzyme *mod* (RM)		Highlander & Hang, 1997
(GTCTC)4	Type I restriction enzyme (RM)	*H. influenzae*	van Belkum *et al.*, 1997a; Appendix 1
(A)7	*p78* (SP)	*Myc. fermentans*	Theiss & Wise, 1997
(A)8	*vaa* (SP)	*Myc. hominis*	Zhang & Wise, 1997
(G)9	*tcpH* (REG)	*V. cholerae*	Carroll *et al.*, 1997
(G)13	*lsi2* (LPS) (=*lgtA*)	*N. gonorrhoeae*	Burch *et al.*, 1997
(G)10	*hpuA* (Fe / SP)	*N. gonorrhoeae*	Chen *et al.*, 1998
(C)11	*lgtG* (LPS)	*N. gonorrhoeae* & *N. meningitidis*	Banerjee *et al.*, 1998; Jennings *et al.*, 1999
(GCAA)n	*nmrep1* (FUN)	*N. meningitidis*	Peak *et al.*, 1999
(GCAA)n	*nmrep2* (FUN/?SP)	*N. meningitidis*	Peak *et al.*, 1999
(GCAA)n	*nmrep3* (FUN/?SP)	*N. meningitidis*	Peak *et al.*, 1999
(CACAG)	Type I restriction enzyme *mod* (RM)	*P. haemolytica*	Ryan & Lo, 1999

SP = surface protein, LPS = gene involved in LPS biosynthesis, REG = regulatory protein, RM = restriction modification system protein, CAP = capsule biosynthesis gene, FUN = function unknown.

That repeats within the coding region directly affect the translation and expression of their associated genes was demonstrated using a *lacZ* fusion with the repeat-containing first gene in the *lic3* locus of *H. influenzae* (Szabo *et al.*, 1992). The influence of the number of tetrameric repeats on expression of LPS phenotypes was demonstrated in *lic2A* by sequencing directly from colonies that had been immunostained with the corresponding antibodies (Maskell *et al.*, 1993; High *et al.*, 1993). Further study of this locus has demonstrated that, within *lic2A*, the SINQ amino acid sequence encoded by the (CAAT)n repeats does not appear to contribute to the function of Lic2A and the CAAT repeat's only apparent function is to confer phase variability upon the reading frame. It was also shown that the number of repeats varies within phenotypic variants of individual strains and between unrelated strains (High *et al.*, 1996). Similar experiments using a *lacZ* fusion of the tetrameric repeat associated *mod* gene in *H. influenzae* initially described by Hood *et al.* (1996) have also demonstrated that the repeat and variation in repeat element length are sufficient to mediate phase variation (de Bolle *et al.*, 1999).

This mechanism of phase variation is generally considered to mediate a clean phenotypic switch such that in the OFF state there is no expression of the associated gene. This has not been formally demonstrated and there is some evidence from the study of *opa* genes cloned into *Esch. coli* that ribosomal frame-shifting in adjacent homopolymeric tracts can facilitate some residual expression of the varied genes (Belland *et al.*, 1989). The biological relevance of this observation or whether it occurs in the normal species contextual background is unknown.

There are several examples that demonstrate that this type of phase variation occurs *in vivo*. Study of the *vaa* genes in *Myc. hominis* which has a homopolymeric tract of As in the 5'region of the ORF (8As is ON, 7 and 9 are OFF) from specimens obtained from an infected joint identified a population with variation in the repeat tract length present during natural infection. However, the observed proportions in this study did not suggest that either has a substantial fitness advantage (Zhang & Wise, 1997). Variation in the LPS of *H. somnus* is known to involve repeat associated genes in a fashion similar to that present in *H. influenzae* (Inzana *et al.*, 1997). Phase variation of these LPS phenotypes was observed to occur in a calf infection model (Inzana *et al.*, 1992). The particular importance of this experiment is that it demonstrates phase variation in a bacterium during long term colonisation of that bacterium's natural host. It also demonstrates that the LPS phenotypes that are present and expressed are immunogenic in the natural host and that there is a selective advantage conferred upon the sub-populations that exhibit phase variation. In human pathogens similar experiments are problematic. In animal model systems variation in population composition has been demonstrated but the extent to which the variation is present in the initial inoculum or occurs during colonisation is difficult to determine (e.g. Weiser *et al.*, 1998a). Evidence that variation in the number of repeats in these genes occurs *in vivo* comes from studies of outbreaks of epidemiologically associated *H. influenzae* infection in which strains that appear to be clonally related when assessed using traditional methods have different lengths of repeat. In this situation, variation in repeat

numbers can be observed in isolates from different individuals (van Belkum *et al.*, 1997a & b). This does not identify when the variation occurs but it does demonstrate that these repeats are unstable and alter the expression phenotype *in vivo*.

The variation of this type of repeat has been consistently demonstrated for many of the loci described in Table 1.3, and also for those located in promoters as described in section 1.5.5. Expression of some of these genes has only been demonstrated to be unstable *in vitro* and in other cases the phenomena have not formally been demonstrated to be reversible (e.g. *tcpH* in VCH (Carroll *et al.*, 1997)). However, it is now reasonable to consider all of the genes in table 1.3 and others which have appropriately located repeats with a base composition and length that are similar to those seen in the above variable genes to be phase variable.

1.5.5 Repeats within promoters

In some genes, the repeats that influence expression are located in the promoter regions and length variation of these repeats affects transcription. The known examples of promoter located repeats that mediate phase variation are summarised in Table 1.4.

Table 1.4. Examples of phase variation mediated by promoter located repeats.

Repeat	Gene	Species	Reference(s)
(C)6	pertussus toxin gene	*B. pertussis*	Locht & Keith, 1986; Nicosia *et al.*, 1986; Gross & Rappuoli, 1989
(C)15	*fim* genes (2, 3 & X) (SP)	*B. pertussis*	Willems *et al.*, 1990
(A)18	*vlp* genes (A, B & C) (SP)	*Myc. hyorrhinis*	Yogev *et al.*, 1991; Citti & Wise, 1995
(TA)10	*hifA – hifB* (SP)	*H. influenzae*	van Ham *et al.*, 1993
(C)13	*opc* (SP)	*N. meningitidis*	Sarkari *et al.*, 1994
(G)11	*porA* (SP)	*N. meningitidis*	van der Ende *et al.*, 1995
(AAT)9	*hxuC* (Fe)	*H. influenzae*	Hood *et al.*, 1996; van Belkum *et al.*, 1997a; Appendix 1

SP = surface protein, Fe = iron metabolism gene. 2, 3, X, A, B and C indicate the specific *fim* and *vlp* genes shown to be phase variable.

The first gene in which a functionally unstable repeat was identified within the promoter was the pertussis toxin gene of *B. pertussis* (Locht & Keith, 1986; Nicosia *et al.*, 1986). It was noted that reduction in the length of a homopolymeric tract of 6 Cs located at the 3′ end of the –35 element affected transcription of the gene (Gross & Rappuoli, 1989). The length of the repeat is sufficient to be consistent with instability and length variation, albeit at the lower end of the range. The spacing within the promoter between –10 and –35 elements is slightly greater than is normally optimal so that alterations in the repeat length are more likely to alter expression. However, the only length variation that has been reported is a reduction to 4 Cs – which is perhaps too short to be reversible at a high frequency and the phenomenon has not been shown to be reversible. It is quite likely that this repeat does mediate phase variation but probably by a different length variation from 6 to 7 Cs; but this remains to be demonstrated experimentally.

The other example in *B. pertussis* is more robust. *B. pertussis* independently phase varies serotypically distinct fimbriae (*fim2* and *fim3*) and also possesses a third homologous gene (*fimX*). Each includes a long homopolymeric tract of Cs that is located 5′ of the –10 promoter component. Variation in the composition of populations with respect to their expression of *fim2* and *fim3* occurs readily in animal models and expression of phenotypes with which the animals have not been vaccinated are favoured in these animals (Preston *et al.*, 1980). These results suggest that this type of variation also occurs *in vivo* in humans in which they are also immunogenic. Three strains which expressed *fim3* had 14 Cs whilst three that did not express *fim3* had 13, 13 and 9 Cs respectively and isolates with variants isolated during infection in a rabbit model showed the corresponding alteration in the length of the repeat element (Willems *et al.*, 1990). Comparison of repeat locations in the *ptx* gene of *B. pertussis* with those of the *fim* genes places the *ptx* associated homopolymeric tract at the same location as the 5′ portion of those in the *fim* genes, further strengthening the case for a functional role of the repeat in *ptx*.

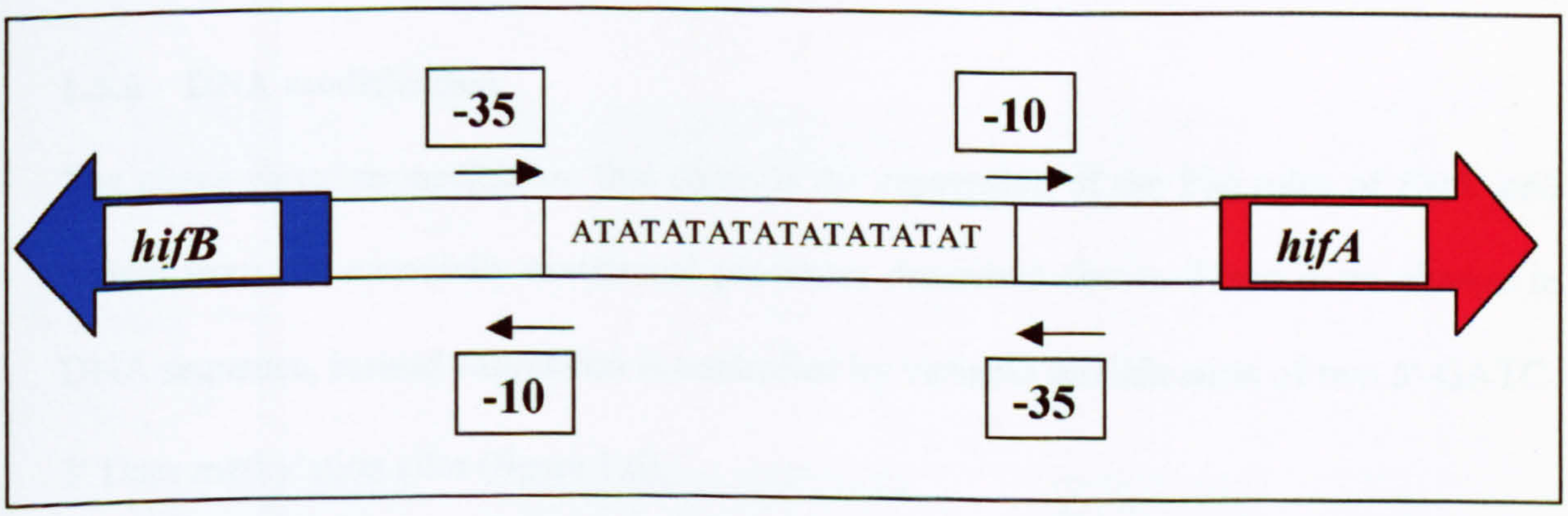
The phase variable *vlp* gene in *Myc. hyorrhinis* has a polyA tract in the equivalent location upstream of the –10 promoter component. Seventeen As results in an ON phenotype, greater than 17 As generates an OFF phenotype. Changes from 17 to 18 As correlated with a change in expression from ON to OFF, and vice versa. It is not really clear whether there are –35 elements in these promoters although they are 'identified' by the authors. Although possible –35 region sequences occur immediately adjacent to the repeat it is not clear whether these or repeated potential accessory binding sites further upstream of the repeat are involved in the control of transcription (Yogev *et al.*, 1991; Citti & Wise, 1995).

The divergent promoters of *hifA* and *hifB* of *Haemophilus influenzae* contain a dinucleotide repeat of (TA)n. This is located between the putative –35 and –10 RNA polymerase binding sites (van Ham *et al.*, 1993) and this affects the transcription of both the fimbrial subunit gene and its chaperone (figure 1.5).

When there are 9 TA repeats the spacing between –35 and –10 units is reduced to 14-bp preventing transcription. When there are 10 TA repeats, the spacing is 16-bp and expression is optimal. Fimbriae are also expressed with 11 and 12 TA repeats present - when the spacing is sub-optimal at 18-bp - and an alternative –35 is probably used when the spacing would otherwise be 20-bp (van Ham *et al.*, 1993; Langermann & Wright, 1990).

The two examples in *N. meningitidis* both have homopolymeric tracts located 5′ of the –10 promoter component. In the *porA* gene the repeat (a homopolymeric tract of Gs) is located immediately adjacent to the –10 (TATAAT) sequence and a sequence (ATGGTT) has been putatively identified to represent the –35 component (van der Ende *et al.*, 1995). In the case of *opc* the repeat, a homopolymeric tract of Cs, is located in the expected location of the –35 (Sarkari *et al.*, 1994). In each case, intermediate phenotypes are observed in addition to ON and OFF phenotypes. Low level expression is also described in the 'OFF' phenotype of the *fim* genes in *B. pertussis* (Willems *et al.*, 1990). Although the most obvious example, *opc* may not be the only case in which there is no –35 consensus

Figure 1.5 Diagram representing the divergent promoters of the flagellar genes *hifA* and *hifB* of *H. influenzae*. The coloured arrows represent the coding regions of the two genes. The black arrows indicate the locations of the promoter components that interact with RNA polymerase. Alterations in the number of copies of the AT dinucleotide repeat illustrated at the center of the diagram alters the distance between the promoter components for both genes simultaneously and mediates phase variation of flagellae. When the optimal number of repeats is present both genes are transcribed. At other repeat lengths transcription is reduced or prevented.



sequence. In each example other that of the *hif* genes of *H. influenzae* and *porA* of *N. meningitidis* the spacing between the –10 and the putative –35 is not optimal, or the putative –35 region is difficult or impossible to identify. In these cases the important promoter components and the mechanism by which the alteration in repeat length alters promoter function has not been defined. Also, the particular characteristics that lead to the generation of intermediate phenotypes by the promoters which contain long homopolymeric tracts of Cs or Gs has not been determined.

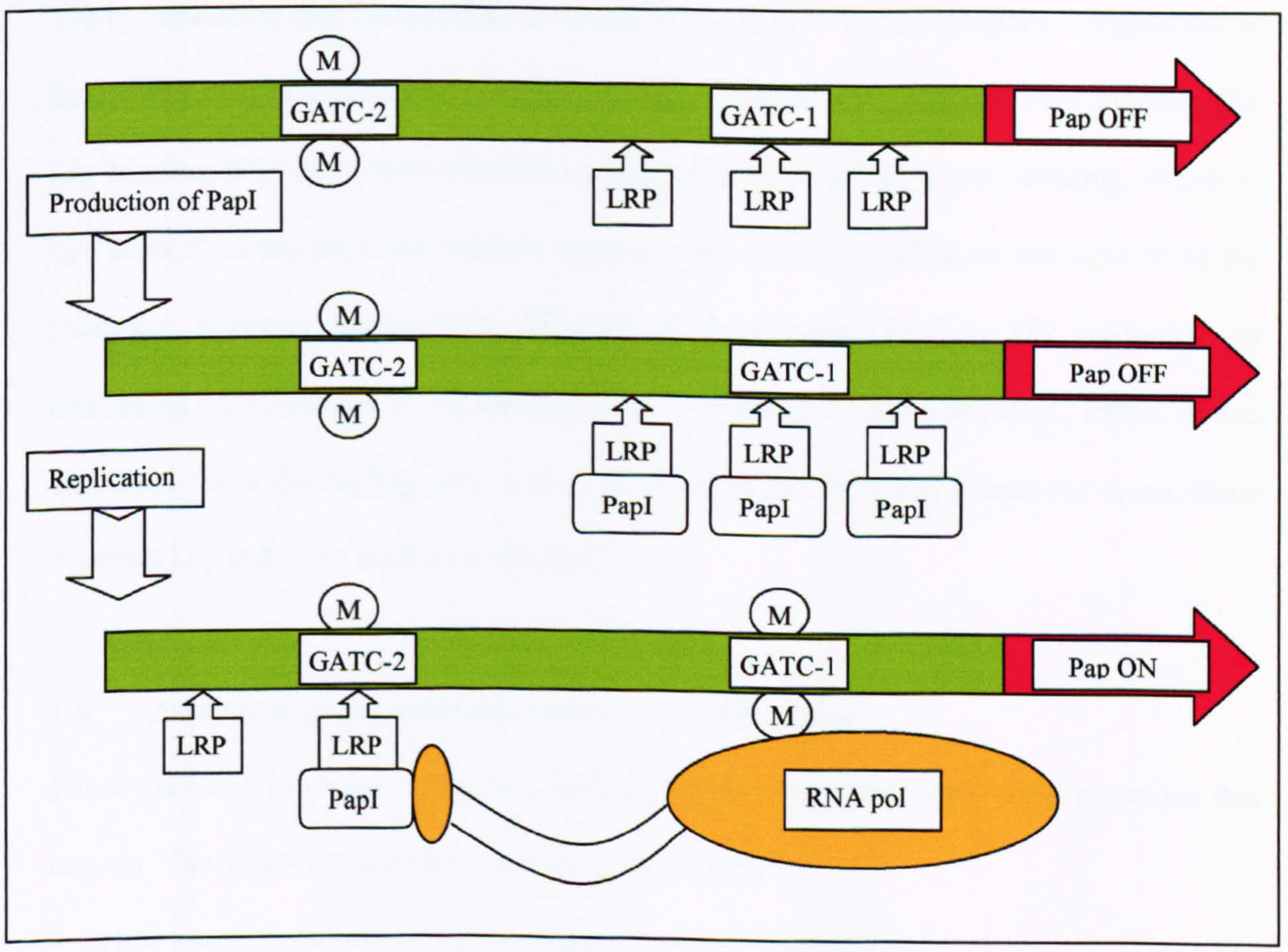
The example from *H. influenzae* of *hxuC* is currently theoretical and based upon genome analysis (Hood *et al.*, 1996; van Belkum *et al.*, 1997a; Appendix 1). There are differences in the number of repeats in the published genome sequence and in the paper originally describing this gene (Cope *et al.*, 1995) and there is variation in the length of the repeat between strains (van Belkum *et al.*, 1997a).

1.5.6 DNA modification

The phase variation mechanism that controls the expression of the Pap pilus of *Esch. coli* differs from the essentially mutational processes described above. There is no change in DNA sequence, instead expression is controlled by variable modification of two 5′-GATC-3′ Dam methylation sites (figure 1.6).

The *pap* operon is composed of at least 11 genes including pilus structural proteins and the proteins necessary for the expression and assembly of the pili. These include two transcriptional regulators: *papB* and *papI*. Two Dam methylation targets are located within the *papI-B* region. GATC-1 and GATC-2 are located 152 and 50-bp upstream of the *papB* transcriptional start site respectively. In phase OFF cells GATC-2 is methylated and GATC-1 is not (figure 1.6), in phase ON cells the situation is reversed. Methylation protection of GATC-2 requires Lrp (but not PapI) (Braaten *et al.*, 1991) suggesting that in phase OFF cells, Lrp binds near to GATC-2 sterically blocking its methylation. Both Lrp and PapI are required for the protection of GATC-1 (Braaten *et al.*, 1991). Footprinting

Figure 1.6 Diagram representing the control of the *pap* gene expression by Dam methylation of the promoter. The mechanism mediating phase variation is described in section 1.5.6.



experiments using Lrp and PapI are consistent with these observations (Nou *et al.*, 1993). That methylation plays a direct role in the control of *pap* transcription is indicated by the absence of transcription in both the presence of excess and absence of Dam (Blyn *et al.*, 1990). Site-directed mutagenesis of GATC-1 to GCTC results in a locked ON phenotype (Braaten *et al.*, 1994). This suggests that methylation of GATC-1 shuts off *pap* transcription and it has been proposed that competition between Lrp and Dam methylase for binding to GATC-1 controls the state of the switch (van der Woude *et al.*, 1992). In addition, the phase locked ON mutant and wild-type bacteria do not express *pap* and alteration of GATC-2 to GCTC results in a phase locked OFF phenotype (Braaten *et al.*, 1994), indicating that methylation of GATC-2 is required for transcription. Regulation of the switch is achieved through a complex interaction between Lrp and PapI binding. Six Lrp binding sites have been identified in the regulatory region of *pap*. Binding of Lrp to two sites 5′ of the promoter inhibits transcription, whereas binding to two sites 3′ of the promoter activates transcription. Binding of PapI favors OFF to ON switching by increasing the efficiency of Lrp binding to the 3′ over the 5′ sites (Nou *et al.*, 1995). In this way PapI shifts the binding of Lrp from GATC-2 to GATC-1 and affects the competition between Lrp and Dam methylase for GATC-1.

1.6 Aspects of phase variation addressed in this thesis

Phase variation has been recognised for a long time but there are still many questions that remain. Those that are addressed in this thesis include:

- How can rates of phase variation be determined accurately?
- How can the influence of phase variation and fitness of the alternate phenotypes on bacterial population structure be modelled?
- What are the complete repertoires of phase-variable genes within bacterial species that use phase variation for which complete genome sequence information is available and what patterns can be detected in phase variable systems?

- How does the promoter located repeat of *opc* function in transcriptional control, particularly how/why does it generate intermediate phenotypes?

PAGE NUMBERING AS ORIGINAL

Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Chemicals

Chemicals were obtained from the following sources;

BDH (Atherstone, Warwickshire), Aldrich (Gillingham, Dorset), Pharmacia Biotech Ltd (St Albans, Herts.) and Sigma Chemical Company (Poole, Dorset) as Analar or Technical grade.

Radiochemicals, [γ^{32}P]dATP and [α^{32}P]dCTP were purchased as "redivue" at an activity of 0.37 Mbq μl^{-1} from Amersham International (Aylesbury, Bucks.). For DNA sequencing [α^{35}S] dATP was obtained from the same source.

2.1.2 DNA modification and other enzymes

Restriction endonucleases were obtained from the following companies: Amersham International, Boehringer Mannheim (Lewes, Sussex) and New England Biolabs (Bishops Stortford, Herts.). Other enzymes were obtained from the following sources:

T4 DNA ligase: Promega (Promega UK, Southampton), Calf intestinal alkaline phosphatase: Boehringer Mannheim, T4 Polynucleotide kinase: Boehringer Mannheim, Klenow: Boehringer Mannheim, Taq DNA polymerase: Boehringer Mannheim, 'Dynazyme' Taq polymerase: FMC (FMC BioProducts, Rockland, ME, USA).

2.1.3 Bacterial growth media and supplements

All bacterial strains were grown at 37°C with 5% CO_2 except where stated otherwise. *N. meningitidis* strains were grown on Levinthals plates consisting of BHI medium (1% agar) supplemented with 5% heated horse-blood. *Esch. coli* strains were grown (in normal

atmosphere) in liquid medium with shaking or on solid L-medium (1% agar). Bacterial strains were stored in growth media with 15-20% glycerol at -80°C.

BHI medium: 3.7% w/v Brain Heart Infusion, Oxoid Ltd (Basingstoke, Hampshire) as per manufacturer's instructions.

L-medium: 1.0% w/v tryptone (Difco, East Molesey, Surrey), 0.5% w/v yeast extract (Oxoid Ltd), and 1.0% NaCl dissolved in distilled water and sterilised by autoclaving (15 p.s.i, 20 minutes). Stored at room temperature.

Media were supplemented with antibiotics for selection of *Esch. coli* or *N. meningitidis* as follows (μg/ml)

	Esch. coli	*N. meningitidis*
Ampicillin	100	n/a
Kanamycin	50	50

Solutions and buffers

Solutions were made by dissolving the ingredients in distilled water. Solutions were stored at room temperature unless stated otherwise.

2.1.4 *Neisseria meningitidis* strains details

MC58 — A virulent strain isolated from an outbreak in Stroud, England. (McGuinness *et al.*, 1991)

MC58 ∝3 — A capsule negative, Opa negative, L3 LPS immunotype, Opc positive, piliated derivative of strain MC58 (Virji *et al.*, 1995).

MC58 ∝11 — A capsule negative, Opa negative, L8 LPS immunotype, Opc negative, non-piliated derivative of strain MC58 (Virji *et al.*, 1995).

N. meningitidis strain obtained from PHLS reference laboratory, Manchester:

Strain	Senders Reference	Phenotype
MN3	88-52570	B4P1.15R
MN11	88-35299	B15P1.7,16R
MN13	87-47626	B14P1.15S
MN19	88-32681	B1
MN26	88-15258	B2a
MN31	89-02264	B4P1.15S
MN32	87-44171	BNT
MN33	87-49513	BNTP1.15S
MN35	87-43701	BNTP1.15R
MN36	88-39017	BNTP1.15S
MN37	88-25268	B2b
MN51	88-13242	B14
MN54	88-12555	B1P1.15S
MN56	89-02359	B4

N. meningitidis strains obtained from the WHO reference laboratory (Caugant *et al.*, 1987):

Strain	ET	Cluster	Phenotype	Source (year)
001	294	E6	B:NT:P1.14	Norway (1982)
BC4	87	A6	B:13	Belgium (1973)
44/76	5	I1	B:15:P1.16	Norway (1976)
P63	198	B4	B:NT:2	Norway (1975)
HF130	104	A10	B:NT	South Africa (1974)
BZ157	?	A4	?	Netherlands (1973)
179/82	228	D4	B:NT:P1.16	Norway (1982)
M470	332	J1	B:NT:-	USSR (?)
M982	133	A16	B:9	USA (1960s)

M986	165	B2	B:2a	USA (?)
S3032	252	D8	B:12	USA (1972)
S3446	8	A1	B:14	USA (1973)

The first strains to be received were those from Manchester. These were received on agar slopes and were grown up on Levinthals plates and used to prepare freezer stocks, DNA using the CTAB method, cell lysates for detection of Opc by immunostaining, and lysates in water for PCR. The lysates in water were sterilised by freezing at –70°C and then boiling for 20 minutes. The WHO strains were received as lyophilised storage vials. These were re-suspended in glycerol broth, an aliquot was frozen and the remainder was cultured and processed as the Manchester strains.

2.2 Preparation of DNA and RNA

2.2.1 Small scale preparation of plasmid DNA

Plasmid containing bacteria (*Esch. coli*) grown overnight in 3 to 5 ml of L-medium supplemented with appropriate antibiotics were pelleted by centrifugation at 10000*g*, 5 min room temperature in a Sorvall Microspin 24S centrifuge and resuspended in 100 µl of TE (25 mM tris-HCl, 10 mM EDTA pH 8.0). The cells were lysed by adding 200 µl of freshly prepared 0.2 M NaOH, 1% (w/v) SDS, mixing by inversion and incubating on ice for 5min. 150 µl of 3 M potassium acetate pH 4.8 was added, mixed gently and left on ice for a further 10 min. This preparation was centrifuged at 10000*g*, 10 min, to remove denatured chromosomal DNA and cell debris. The supernatant was removed and the DNA precipitated by addition of 2 volumes of ethanol. The plasmid DNA was pelleted by centrifugation (10,000*g*, 15 min, 4°C), and washed twice with 70% ethanol before air drying, and final resuspension in 50 µl of TE containing 10 µg ml^{-1} RNase A.

2.2.2 Large scale preparation of plasmid DNA

Plasmid containing bacteria were grown overnight in 25 ml of L-broth supplemented with appropriate antibiotics and plasmid DNA purified using a Qiagen-tip 100 according to manufacturer's instructions and using reagents supplied (Qiagen, GmbH).

2.2.3 Preparation of Bacterial Genomic DNA

Genomic DNA was prepared from using selective precipitation with CTAB to remove cellular proteins and cell wall components. Bacteria were grown overnight on solid media. Approximately 1/2 a plate of semi-confluent growth was harvested into 1.13 ml TE buffer, and the cells lysed by the addition of 60 µl of 10% (w/v) SDS and 6 µl of 20 mg ml^{-1} proteinase K and incubated at 37°C for 1 hour. 200µl of 5M NaCl was added and mixed thoroughly. 160 µl of CTAB/NaCl solution was added, mixed and then incubated at 65°C for 10 min. An approximately equal volume of chloroform was added, mixed and centrifuged at 10000g for 5 min. The supernatant was removed and the chloroform extraction repeated. DNA was precipitated by addition of 0.6 volumes of isopropanol, and removed with a sterile plastic loop. DNA was washed twice with 70% ethanol, before air drying and final resuspension in 200 µl TE containing 10 µg ml^{-1} RNase A.

CTAB/NaCl solution; 10% hexadecylcetyltrimethylammonium bromide (CTAB) in 0.7 M NaCl.

2.2.4 Quantitation of DNA

Concentrations of DNA solutions were estimated by two methods. Visual comparison of sample DNA to standards of known concentration by visualisation by UV transillumination after agarose gel electrophoresis gave approximate concentrations. For more accurate estimation of DNA concentration samples were examined in a spectrophotometer at a wavelength of 260 nm. Concentrations were estimated assuming that 1.0 OD unit represents 50 µg ml^{-1} for double stranded DNA and 30 µg ml^{-1} for oligonucleotides.

2.3 Modification of DNA

2.3.1 Restriction endonuclease digestion

DNA was incubated with restriction endonucleases in buffers supplied by manufacturers, and at the recommended temperatures. For restriction mapping of plasmids, approximately 500 ng of DNA was digested. For purification of plasmid fragments after electrophoretic separation, approximately 2 μg of DNA was digested. For Southern blot analysis of genomic DNA, approximately 10 μg total DNA was digested. For purification for construction of an enriched library, 20 μg of chromosomal DNA was digested.

2.3.2 DNA ligation

Ligation reactions were performed by addition of 2U of T4 DNA ligase in buffer supplied by the manufacturer. Reactions were carried out overnight at 15°C.

2.3.3 Klenow reactions

5′ protruding termini resulting from restriction digestion of DNA were filled by incubation of DNA with Klenow enzyme. Reactions were carried out at 37°C in the presence of 2.5 mM dNTPs in the buffer used for restriction digestion.

2.4 Agarose gel electrophoresis of DNA

DNA gel electrophoresis was conducted in a horizontal mini-subcell at 100 V or sub-cell equipment (BioRad Laboratories, Hercules, CA) at 20V overnight in 1 X TBE buffer. Agarose gels were prepared with electrophoresis grade agarose dissolved in TBE and contained ethidium bromide (EtBr) at a final concentration of 1 μg ml^{-1}. Percentages of agarose varied from 0.8% for DNA fragments >1 kilobase (kb) to 2% for resolution of fragments <1 kb. For resolution of fragments smaller than 400 bp gels were prepared using

Metaphor® (FMC BioProducts, Rockland, USA) according to the manufacturers instructions. An appropriate volume of loading buffer was added to each DNA sample before loading. DNA bands were visualised under ultraviolet (UV) transillumination and photographed with a Polaroid Cu.5 land camera using Polaroid type 665 or 667 film. Sizes of DNA fragments were determined by comparison with migration of markers of known molecular weight.

TBE buffer (5 X); Tris base, 0.445 M, Boric acid, 0.445 M, EDTA (pH 8.0) 10 mM

6 X DNA gel loading dye; Glycerol, 30% (v/v), Bromophenol blue, 0.25% (w/v), Xylene cyanol, 0.25% (w/v) in TE buffer

2.5 Purification of DNA fragments from agarose gels

DNA bands were visualised under UV transillumination and the fragments of interest excised using a sterile scalpel blade. DNA was purified using the Qiaex II gel extraction kit according to the manufacturer's instructions and using reagents supplied.

2.6 Polymerase Chain Reaction (PCR)

2.6.1 Templates

Three different types of DNA sample were used as PCR templates:

i) genomic DNA. Between 1-10 ng of genomic DNA was used as template for PCR, prepared as described in section 2.2.3.

ii) plasmid DNA. Less than 10 ng of plasmid DNA was used as template. Typically estimated as 0.5 µl of a 1 in 1,000 dilution of a plasmid preparation.

iii) whole cell lysates. Approximately 10 colonies were suspended in 100 µl sterile water and subjected to two cycles of freeze-thaw. 2 µl of this was used as template for PCR.

2.6.2 PCR conditions

Each PCR contained the appropriate template, 20 ng of each primer, 1.6µl of 25 mM dNTPs, 5 µl 10 X PCR buffer and 2 U of Taq DNA polymerase. Reaction volumes were made up to 50 µl with sterile distilled water, and overlaid with mineral oil. PCR was performed with *Taq* DNA polymerase (Promega or Boehringer) under the conditions recommended by the supplier. All the PCR experiments included a negative control i.e. a reaction containing all the reagents but without the template DNA. The thermal cycling was conducted in a Perkin Elmer DNA Thermal Cycler. A 30 cycle program was typically used- [30 secs at denaturing temperature, 94°C; 30 secs at annealing temperature, 50-61°C; 1 min at elongation temperature, 72°C]. Annealing temperature was chosen as 2°C below the calculated annealing temperature for the primers where T (°C)= 2(T+A) + 4(G+C). Following amplification a 5 µl aliquot of the reaction was checked on an agarose gel.

2.6.3 Direct sequencing of PCR products

One primer used in the PCR was purchased 5′ end labelled with biotin and used under standard reaction conditions. The resulting product was purified using Dynabeads (Dynal) according to manufacturer's instructions, before sequencing using the Sequenase kit (USB) (section 2.10).

2.6.4 Cloning of PCR products

PCR products were cloned into the PCR cloning vectors pCRII (Invitrogen) or pT7Blue (Novagen) according to manufacturer's instructions.

2.7 Southern blot analysis

2.7.1 Labelling of DNA probes

2.7.1.1 Radiolabelling

Double stranded DNA was radiolabelled with [α^{32}P]dCTP using the Megaprime Kit (Amersham) according to the manufacturer's instructions and using the reagents supplied. Oligonuclotides were labelled by incubating 2 pmol of DNA with 3 pmol [γ^{32}P]dATP and 1 unit T4 polynucleotide kinase at 37°C for 20 minutes. Unincorporated dNTPs were separated from labelled DNA fragments by passage of the labelling reaction mixtures through Nick (G50) or Nap (G25) Sephadex columns for double stranded and oligonucleotides respectively (Pharmacia).

2.7.1.2 Digoxigenin labelling

Probe DNA was prepared by PCR amplification of the *opc* open reading frame using primers Opc1 and Opc2. This was gel purified and extracted as described in section 2.5. Approximately 0.5 μg of extracted DNA was labelled using the DIG DNA labelling kit (Boehringer Mannheim, GmbH, Germany) according to the manufacturer's instructions. Half of the labelled material was used in a single hybridisation.

2.7.2 Electrophoresis, Southern blotting, hybridisation and filter washing

DNA was separated by gel electrophoresis as described in section 2.5 then denatured and neutralised as described previously (Sambrook *et al.*, 1989). After overnight blotting, DNA was fixed to the nylon membrane by exposure to UV illumination for 2 min.

Denaturation solution: 1.5 M NaCl, 0.5 M NaOH.

Neutralisation solution: 0.5 M Tris-HCl pH8, 1.5M NaCl.

SSC: 150 mM NaCl, 15 mM sodium citrate.

Blots were incubated at hybridisation temperature in hybridisation solution for 2 h prior to addition of radio-labelled probe. Unless otherwise stated, hybridisation was carried out for 16 hours at 45°C and 65°C for oligonucleotide and double stranded probes respectively, followed by successive 15 minute washes in (6 X SSC, 0.1% SDS), (2 X SSC, 0.1% SDS), (1 X SSC, 0.1% SDS), (0.5 X SSC, 0.1% SDS), and (0.2% X SSC, 0.1% SDS) at the hybridisation temperature.

Hybridisation solution: 6 X SSC, 5 X Denhardt's solution, 0.5% sodium dodecylsulphate (SDS), 100 μg/ml sheared salmon sperm DNA.

50 x Denhardt's solution: 1% w/v ficoll, 1% w/v polyvinylpyrrolidine, 1% w/v bovine serum albumin (BSA). Stored at -20°C

2.8 Transformation of bacteria

2.8.1 Preparation of competent *Esch. coli* cells

An overnight culture of *Esch. coli* was inoculated into 10ml of growth medium at 1:100 dilution and grown to OD600=0.4-0.6. The culture was then chilled on ice for 20 min. The cells were pelleted by centrifugation at 3000g at 4°C for 10 min. The cell pellet was resuspened in 10ml ice cold 0.1M $CaCl_2$ and incubated for 10 min on ice. The cells were repelleted as above and the pellet resuspended in 5ml 0.1M $CaCl_2$ and incubated for 10 minutes on ice. This step was then repeated and the cells were then incubated on ice for between 1-12 hours before use.

2.8.2 Transformation of *Esch. coli*

200 μl of competent cells were added to an incubated ligation reaction. This mixture was incubated on ice for 30 to 60 minutes before transfer to 42°C for 2 minutes. The mixture was then placed back on ice for 30 seconds. 1ml of growth media (without antibiotic) was added and then incubated at 37°C for 30 minutes before plating onto selective media.

2.8.3 Transformation of *N. meningitidis*

4-5 colonies from overnight growth were pooled and spotted onto fresh solid media and regrown for 30 minutes. Plasmid constructs for transformation were linearised by restriction digestion prior to transformation. 10 µl of these restriction digests were spotted onto the bacteria, and incubated for 5 hours to allow uptake of DNA. The resulting bacterial growth was scraped from the media using a sterile microbiological loop and streaked onto selective media.

2.9 Screening of λ-Zap II library

A λ-Zap II library had been previously prepared using genomic DNA of *N. meningitidis* strain MC58 (Jennings *et al.*, 1995). This was screened according to the manufacturer's instructions. The library was plated out to approximately 100,000 pfu per plate (23 cm x 23 cm) and transferred to nylon membranes prior to hybridisation with double stranded DNA probes. The PCR generated probes, generated with primers Opc1 and Opc2 were prepared as described in section 2.7.1.1. Plaques giving hybridisation signals were excised and the screening procedure was repeated at a plate density of 1000 pfu to obtain well-isolated plaques. The positive plaques were excised and transformed into the *Esch. coli* strain provided according to the manufacturers instructions.

2.10 DNA sequencing

2.10.1 Sequencing reactions

Double stranded DNA was sequenced by the dideoxynucleotide chain termination method (Sanger *et al.*, 1977) from plasmid preparations. DNA was denatured by addition of 2 µl 1N NaOH to 10µl DNA and incubation at 37°C for 30 min, before neutralisation with 3 µl 3M NaOAc (pH 5.5). After ethanol precipitation, subsequent steps were performed using a Sequenase version 2.0 DNA sequencing kit (U.S.B., Cleveland, OH) and the protocol

supplied with the kit. Primers were used at 10pmol/reaction with approximately 1μg of DNA. Standard conditions involved a 5 min, room temperature incubation with the Sequenase enzyme and a 1/5 dilution of the labelling mix followed by 5 min, 37°C termination reactions. To obtain sequence close to the primer a 1/10 dilution of the labelling mix and incubation times of only 3 min, were used while longer sequences could be read with undiluted labelling mix.

2.10.2 Sequencing electrophoresis

Sequencing reactions (4μl) were loaded onto a 6% acrylamide:bisacrylamide (19:1) gel containing 8% (w/v) urea as denaturant using solutions from the Sequagel kit (U.S.B., Cleveland, OH, USA). Samples were heated to 75°C for 5 min and placed on ice to denature the DNA prior to loading on the gel which had been pre-warmed to approximately 45°C by running at 60 W for 1hr. Electrophoresis continued for 2-6hr depending on the distance from the primer to be read. The sequencing gel was fixed in a solution containing 10% methanol (V/V) and 10% acetic acid (v/v) for 15 minutes, then transferred to 3MM paper, covered in Saranwrap, dried for 20 to 40 mins at 80°C in a vacuum gel drier (BioRad) and then exposed to autoradiographic film at room temperature. Sequence was read manually before compilation and analysis using the GCG software package (Genetics_Computer_Group, 1996).

2.11 SDS-PAGE

Proteins were analysed by SDS-PAGE following the method of Laemmli (Laemmli, 1970) using a large format Biorad gel system (Protean II model). 12.5% (w/v) polyacrylamide (38:1) gels were run. Protein samples were boiled for 5min in 1-2 volumes of 2X sample buffer before loading on the gel. Large gels were electrophoresed overnight in tank buffer at 15-18mA constant current. Duplicate gels were run and then either stained, or Western blotted. Gels were stained in Coommassie blue R250 or silver-stained using the Quick-

Silver kit (Amersham) as per manufacturer's instructions. Molecular weights of proteins were determined using Rainbow Mwt (Amersham) or Sigma silver stain protein markers according to the suppliers instructions.

Sample buffer (2 X): Tris-HCl (pH 6.8) 125 mM, Glycerol 4% (v/v), SDS 4.6% (w/v), Bromophenol blue 0.005% (w/v), 2-mercaptoethanol 5% (v/v).

Stacking-gel buffer: Tris-HCl (pH 6.8) 0.5 M,SDS 0.4% (w/v).

Separating-gel buffer: Tris-HCl (pH 8.8) 1.5 M, SDS 0.4% (w/v).

Tank buffer: Tris base 125 mM, Glycine 0.96 M, SDS 0.5% (w/v) pH 8.3

2.12 Tract length determination by direct Southern blotting

This was done using the protocol previously described by Jennings *et al.* (1995). Samples of chromosomal or plasmid DNA were digested with *Tsp*509I to generate a fragment of 52 bases including the homopolymeric tract for a repeat length of 12 bp. 0.66 volumes of Sequenase stop mix (USB) was added to the digested DNA, denatured at 65°C for 5 minutes, and then run on a 6% denaturing polyacrylamide sequencing gel. The plates were then separated and a piece of Hybond-N membrane (Amersham) was laid directly onto the gel. This was covered by a piece of Whatman 3M filter paper of the same size, then by a glass plate and weights, and left to blot for 10 minutes. The plate and filter paper were then removed and any bubbles underneath the now wet membrane were then removed by rolling with a 10 ml pipette, and the blotting repeated with fresh filter paper for a further 30 minutes. The filter was then removed, air dried, UV fixed, and pre-washed in 6 x SSC and then hybridised using a ^{32}P radiolabelled oligonucleotide as described in section 2.7.1.1.

2.13 TBE-PAGE

TBE-PAGE gels were used for the separation of DNA restriction fragments prior to silver staining and also for EMSAs. A typical 5% gel used for EMSA was prepared using 8.75ml of 20% 29:1 acrylamide/bisacrylamide stock solution (Sigma), 1.75 ml of 10 x TBE, 24.5

ml of water, 300 µl of 10% APS and 75 µl of TEMED. The gel concentration was increased up to 15% when separating DNA fragments for silver staining by increasing the proportion of acrylamide/bisacrylamide and decreasing the amount of water accordingly.

2.14 Silver staining of DNA

Silver staining of restriction digest DNA products was performed using a modification of the method described as modification 1 of Johnasson and Skoog described by Vari and Bell (1996). The final protocol, using Milli-Q quality water throughout, was:

The gel was washed in water 3 times for 10 minutes.

The DNA was sensitised with 12% glutaraldehyde for 30 minutes.

The gel was washed in water a minimum of 6 times for 10 minutes (may be increased to 10 times if time permitting).

DNA was bound to the silver using a freshly prepared staining solution of 8 g/l $AgNO_3$, 10 ml/l 25% NH_3, 50 ml/l 1M NaOH (adding the NaOH prior to the NH_3) for 30 minutes.

The gel was washed in water 3 times for 10 minutes.

Silver staining was developed using a freshly prepared solution of 500 µl/l of HCHO and 0.05 g/L of citric acid for up to 30 minutes depending upon the rate of band stain development.

The staining reaction was stopped using a 5% solution of acetic acid for 5 minutes.

2.15 Labelling of PCR slippage products

PCR reactions were performed as described in section 2.6. Amplified products were ethanol precipitated by adding $1/10^{th}$ volume of 3M sodium acetate solution and 3 volumes of ethanol, and processed similarly to plasmid preparations as described in section 2.2.1. The DNA was resuspended in restriction digest buffer and digested for 2 hours with *Hin*dIII. The digested PCR product was run out on a 3% Metaphor gel and the fragment was excised and purified as described in section 2.5. These were then end-labelled using a reaction of 8 µl

(of 20) of Quiex extracted DNA, 1 µl restriction buffer L, 1 µl of a 1 in 5 dilution of ^{35}S dATP, and 0.5 µl of Klenow (1.5 enzyme units). This was incubated for 15 minutes at room temperature and the reaction was terminated by the addition of 6.5 µl of STOP solution from the Sequenase sequencing kit. This was then resolved on sequencing gels as described in section 2.10.2. Radiolabelled bands on the gels were either detected using photographic film or a PhosporImager® screen (Kodak / Molecular Dynamics, USA) and read using a Storm 840 PhorsphorImager® (Molecular Dynamics, USA). Phosphorimager data files were processed and analysed using ImageQuant® software (Molecular Dynamics, USA).

2.16 Extraction of repeat containing restriction fragments

Restriction digests mixed with the oligonucleotide were mixed and denatured at 95°C for 5 minutes. The digested fragments and the oligonucleotide were allowed to anneal by slow cooling to 5°C below the calculated annealing temperature of the oligonucleotide over 30 minutes in a thermal cycler. The repeat containing DNA fragment was then purified on Dynabeads using the protocol used for template preparation used in the sequencing of PCR products (section 2.6.3) and loaded onto denaturing PAGE gels with the denaturing 'STOP' solution from the Sequenase® sequencing kit.

2.17 Immunological detection of Opc expression

Opc expression was determined by colony immunoblotting and immunostaining of serially diluted cell lysates. Cell lifts were lifted, or 10 µl of serially diluted cell lysates were placed onto nitrocellulose discs. These were blocked with 5% skimmed milk in PBS containing 0.05% Tween-20 (PBST) and immunostained with a 1:2000 dilution of Opc specific antibody B306 (Achtman *et al.*, 1988) for 30 minutes. Immunoblots were washed three times with PBST and then reactivity was detected using rabbit anti-mouse antibodies

conjugated to alkaline phosphatase (Sigma) and nitroblue tetrazolium and 5-bromo-4-chloro-3-indolylphosphate (Sigma) as substrates.

2.18 DNA binding protein preparation

2.18.1 Polyethylenamine precipitation

Cells were harvested from Levinthals plates into PBS on ice and then freeze thawed twice at –70°C to kill the bacteria. The bacteria were then pelleted by centrifugation at 1500 rpm (1000g) in a bench centrifuge and re-suspended in TGED (TGED: 0.01M TRIS pH 7.9, 0.5M EDTA, 25% volume glycerol, 1mM DTT) adjusted to 0.2M NaCl. DNA was sheered ultrasonically using a Sonifer 250 ultrasonicator fitted with a Microprobe (Branson, Danbury, CT, USA) with 60% power output and 60% pulse cycles for 5 bursts of 30 seconds. Cell debris was pelleted by centrifugation at 15,000 rpm (30,000g) for 30 minutes in a Beckman J2-21 centrifuge (Beckman Instruments, Beckman Coulter, Fullerton, CA, USA). The supernatant was removed and 10% polyethylenamine (polymin-P) solution in TGED was added drop-wise with good mixing to a total of 35 µl per ml of lysate, at which point a precipitate formed and there was visible clearing of the cell extract. The precipitate was then washed once with 0.2M NaCl TGED to remove un-complexed proteins from the pellet. The precipitate was then washed once with 0.5 M NaCl TGED to elute a large proportion of the low molecular weight proteins. The precipitate was then washed again with 1.0 M TGED to elute the high molecular weight proteins presumed to include RNA polymerase and the supernatant was retained. If the protein was not to be used immediately then the glycerol was increased to 20% v/v and the protein extracts were frozen at -80°C.

2.18.2 Ammonium sulphate precipitation

Saturated ammonium sulphate solution was prepared by dissolving 900g of ammonium sulphate in a total volume of 1 l of water, on a heated stirring plate, until completely dissolved. The solution was filtered and the pH was adjusted to 7.6 using ammonium

hydroxide. This was added to protein extracts such that the final concentrations were increased by increments of 10%, or other intervals as described (Englard & Seifter, 1990).

2.18.3 Desalting of protein extracts

Protein extracts were desalted using HiTrap minicolumn (Amersham Pharmacia Biotech AB, Uppsala, Sweden) according to the manufacturers' instructions.

2.18.4 Heparin affinity chromatography

Heparin Econo-Pac Cartridge mini-columns containing 5 ml of Affi-Prep heparin support (Bio-Rad Laboratories, Hercules, CA, USA) were used according to the manufacturers instructions using 0.15 M and 2 M NaCl TGED as buffers 1 and 2.

2.18.5 Polyethylene glycol precipitation

This method consists of the first 4 steps of the method of Gross *et al.* (1976) for the purification of RNA polymerase. This method uses PEG precipitation under low salt conditions to precipitate DNA and DNA bound proteins, and subsequent elution of these proteins under high salt conditions. This uses the following solutions:

TGED: 0.01 M TRIS pH 7.9, 0.5 M EDTA, 25% v/v glycerol, 1 mM DTT

Solution A: 0.01 M TRIS, pH 7.9, 25% w/v sucrose, 0.1 M NaCl

Solution B: 0.3 M Tris pH 7.9, 0.1 M EDTA, 4 mg per ml of lysosyme (added just before use)

Solution C: 1.0 M NaCl, 0.02 M EDTA, and 0.08% w/v deoxycholate

Solution D: 17% w/v polyethylene glycol 6000, 0.157 M NaCl, 0.01 M DTT (added just before use)

Solution E: 5% w/v polyethylene glycol 6000, 2 M NaCl, 0.01 M TRIS pH 7.9, 0.01 M DTT (added just before use)

Cells from 10 semi-confluent 10 cm Levinthals plates were harvested and suspended in PBS and pelleted by centrifugation at 6000 rpm (4360g) in a Beckman J2-21 centrifuge (Beckman Instruments, Beckman Coulter, Fullerton, CA, USA). The pellet was suspended and lysed by the sequential addition of 1 ml of solution A, 0.25 ml of solution B, and 1.25 ml of solution C. The DNA and bound proteins were precipitated by the addition of 3.5 ml of solution D and the mixture was thoroughly mixed by vortexing. The precipitate was pelleted by centrifugation at 7000 rpm (5930g) and the supernatant was discarded. 0.5 ml of solution E was added to the pellet and this was gently re-suspended using a plastic disposable colony picker to elute the DNA binding proteins. The precipitate was pelleted by centrifugation at 8000 rpm (7740g) and the supernatant was collected. The material was then either diluted with 6.15 ml of TGED to reduce the NaCl concentration to 0.15M, or the material was desalted using a HiTrap minicolumn (Amersham Pharmacia Biotech AB, Uppsala, Sweden) according to the manufacturers' instructions prior to running on SDS-PAGE gels.

2.18.6 Affinity purification using the *opc* promoter

50µl of streptavidin coated Dynabeads were washed once with 100 µl of TNE and added to 90µl of promoter region PCR product amplified using OpcPro and Opc16 using the appropriate cloned and sequenced promoter region in pCRII as template. The DNA bound to the beads was washed once with 100 µl of TNE and once with 100 µl of TGED. A DNA binding reaction was composed of 187µl of TGED, 2.5 µl of double stranded poly(dI-dC).poly(dI-dC) and 20 µl of the protein extract prepared as described in section 2.18.5. This was incubated for 30 minutes at room temperature and the supernatant was then removed. The beads were washed once with 100 µl of TGED and then the promoter bound proteins were eluted using TGED with 2M NaCl. Proteins in the eluate were resolved on an 8% SDS-PAGE gel (with a 4% stacking gel) using a ratio of 37:1 acrylamide to bisacrylamide (Sigma).

2.19 Electrophoretic mobility shift assays

This was done using standard methodology (Hennighausen & Lubon, 1985). Template DNA was initially generated by excision of cloned promoter regions in pCRII (amplified using OpcPro and Opc16) and end-labelling by end-filling using α^{32}P dCTP (section 2.3.3). This method often resulted in several artifactual bands on EMSA for unknown reasons. Later experiments used PCR amplified promoter regions from the using the same plasmids as templates, using OpcPro and Opc16, with a γ^{32}P end-labelled Opc16 primer.

Reaction conditions for end filling DNA excised from plasmids were: 28 µl of water, 4 µl (about 500 ng) of gel purified DNA, 10 µl of Klenow buffer, 5 µl 100 mM dGTP solution, 1 µl acetylated BSA, 1 µl of α^{32}P (0.37 MBq) and 1 µl of Klenow (3 units). The labelled DNA was prepared for protein binding by cleaning phenol-chloroform, ethanol precipitation, and re-suspension in water.

Oligonucleotides were labelled using polynucleotide kinase (PNK). Reaction conditions for oligonucleotide: 50 pMol of oligonucleotide in 1µl was incubated with 1µl of water, 1µl of PNK buffer (supplied with the enzyme), 1 µl of PNK and 3 µl of 2.5 µM γ^{32}P dATP stock solution. This would contain in the region of 1 MBq of radioactivity.

Final reaction buffer conditions for DNA-protein binding were 10mM TRIS pH 7.5, 5mM MgCl, 2mM DTT, 0.12% Triton X-100, 0.1 mg/ml BSA and 0.05mM EDTA. To this poly dI-dC was titrated until specific binding was obtained using the method of Hennighausen & Lubon (1985). The optimal amount of poly(dI-dC).poly(dI-dC) was found to be 2.5µl of a 2 mg/ml solution in a final volume of 20 µl. Reactions were assembled including all ingredients except the template DNA and incubated at room temperature for 15 to 20 minutes. Radiolabelled DNA (approximately 10,000 counts) was then added and the reaction was incubated for a further 20 to 30 minutes. At the end of the binding incubation 1/10 volume of loading buffer was added and the reaction was loaded directly onto a TBE-

PAGE gel and run at low voltage (15 to 20V) overnight at 15°C (using a thermostatically controlled water cooling system attached with the BioRad Protean II gel cooling unit).

2.20 Site directed mutagenesis strategies

Two methods were used for site directed mutagenesis.

The homopolymeric tracts were replaced using the strategy illustrated in figure 2.1. The repeat region was changed using an oligonucleotide that extended from the adjacent *Hin*dIII site, through the entire length of the homopolymeric tract location including the engineered changes, and with 26 bases adjacent to the repeat that were sufficient for priming the PCR. A separate primer was obtained for each replacement sequence that was used (section 2.23.4). Each primer was used with LT7 (section 2.23.1) to amplify the 5′ 780 bp of the insert containing *opc* in pNJS1 and the pBluescript polylinker up to and including the T7 promoter. Each PCR product was cloned into pCRII. The entire cloned region containing the altered *opc* promoter region was excised using the *Hin*dIII site adjacent to the replaced homopolymeric tract and the *Kpn*I site in the cloned pBluescript polylinker. These were cloned into the equivalent sequence locations of pNJS1 as described in section 7.6.

The strategy used to alter the IHF site is illustrated in figure 2.2. The putative IHF binding site of the site directed mutant of pNJS1 that contained the replacement tract equivalent to 13 Cs (pC13R) was mutated using an overlapping PCR method. The overlapping oligonucleotides IHF-F and IHF-R (section 2.23.6) were used to amplify the 3′ portion of the pC13R insert (with LReverse) and the 5′ portion portion of the pC13R insert (with LT7). The PCR products from these two reactions were mixed and used as a template in a further PCR using LT7 and LReverse primers to amplify the whole of the previous insert which now contained the altered IHF site in addition to the sequence replacing the homopolymeric tract and the kanamycin cassette. The whole PCR product was then used to transform *N. meningitidis* strain MC58 ∝3 and transformants were selected on kanamycin. Kanamycin resistant transformants were screened with IHF-F and Opc2, and OpcPro and RepScreen to

Figure 2.1 Showing the process of site directed mutagenesis of the *opc* promoter homopolymeric tract

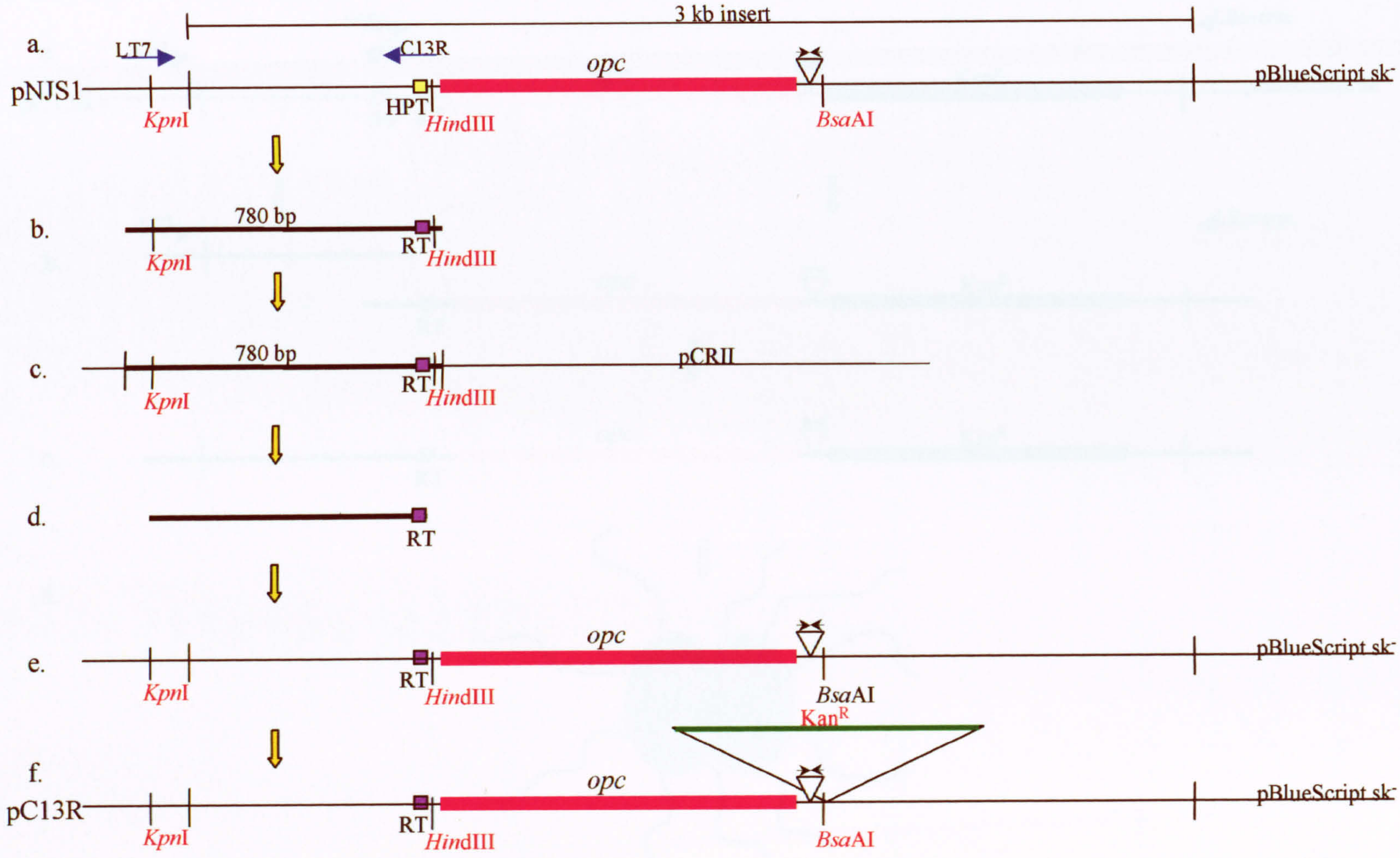
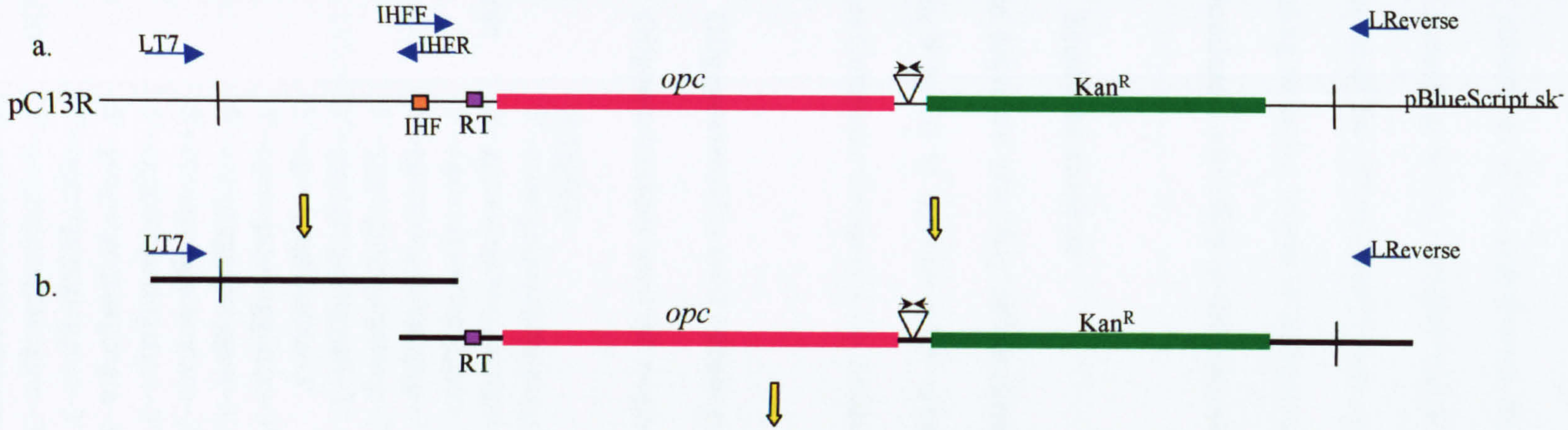


Figure 2.2 Showing the mutagenesis of the IHF consensus binding site in the *opc* promoter



ensure that both new IHF site and the previous homopolymeric tract replacement sequences were present.

2.21 Cloning and expression of the α subunit of RNA polymerase

The α subunit of RNA polymerase from *N. meningitidis* was cloned into the pRSET T7 vector according to the manufacturer's instructions (Invitrogen corp, Carlsbad, CA, USA). Expression of the cloned protein was induced using IPTG and was purified from cell lysates by binding to nickel bound to an agarose support (Ni-NTA, Qiagen, Gmbh) and was eluted using imidazole according to the manufacturer's instructions.

2.22 Sequence analysis

Routine sequence assembly and analysis was performed using the GCG software package (version 6 to 10 as updates were released) (Wisconsin Package Versions (6.0 to 10.0), Genetics Computer Group (GCG), Madison, WI, USA).

2.23 Oligonucleotides used in this project

2.23.1 Oligonucleotides used for sequencing:

primer	sequence	position in sequence (Fig 7.1)	
LT7	5′–cattatgctgagtgatatcccgct–3′	vector	
LReverse	5′–ggaaacagctatgaccatgat–3′	vector	
SK	5′–cgctctagaactagtggatc–3′	vector	
Opc 1	5′–ggaagccggtttcgggcg–3′	1041-1058	
Opc 2	5′–gggatggtgagcgatacg–3′	1598–1581	
Opc 3	5′–gggtgccggcttgggtt–3′	1215–1230	
Opc 4	5′–cgccggtgtttaattta–3′	1408–1392	
Opc5	5′–cgcccgaaaccggcttcc–3′	1921–1932	
Opc6	5′–ccgtgtgggtgccggctt–3′	1209–1226	
Opc7	5′–cccaagccggcacccaca–3′	1229–1212	
Opc8	5′–cggccccgcctcgatgct–3′	1518–1501	
Opc10	5′–gccgacgcgcaagccgta–3′	1635–1618	
Opc11	5′–cggcttgcgcgtcggcat–3′	1620–1637	
Opc12/20	5′–gcggcggcagtaccggtc–3′	884-867	b
Opc13	5′–gaaaaccgaaatcagaaacc–3′	377–396	
Opc14	5′–attctcacttggttttctgtt–3′	671–651	
Opc15	5′–taacaattcgttgtaaacaaata–3′	671–693	

Opc16	5′–gtagtcggatatggtaacat–3′	803-784	b
Opc17	5′–caatccaaattttggagatttt–3′	803–824	
Opc18	5′–gcaatcatggcacatgtaaaa–3′	863–843	
Opc19/21	5′–cccgctttcgcggaatgacg–3′	291-311	b
Opc22	5′–gccgccatctcaagtctcgtcattccct–3′	342-369	b
Opc23	5′–attttacatatattaataaaaattaacaaa–3′	727–698	
B28-Seq1	5′–catatattaataaaaattaac–3′	721–701	
B28-Seq2	5′–cggtgtgtggatttatgcg–3′	1879–1897	
B28-Seq3	5′–ccggcataaataccgcttt–3′	3′ of *opc*	
B28-Seq4	5′–cgggtatccggggaggat–3′	58–75	
B28-seq5	5′–gccgccatctcaagtctcgtcattccct–3′	342–369	

2.23.2 Oligonucleotides used to clone the S3446 *opc* sequence region:

Opc25	5′–cgggggaggattaaggggtatttggg–3′	65–91	
Opc26	5′–ggccggtttcgcaaaaaacaatcaaaa–3′	1931–1905	

2.23.3 Oligonucleotide used to probe Southern blots:

HPT-Probe	5′–atattcttaagctttcggggg–3′	755-735	b

2.23.4 Oligonucleotides used for site directed mutagenesis of the homopolymeric tract:

All start and finish (752-702) in the same position and include the sequence replacing the homopolymeric tract.

C10R	5′–ttcttaagctttcgccgcgccgcattttacatatattaataaaaattaa–3′		
C11R	5′–ttcttaagctttcgccgcgccgcgattttacatatattaataaaaattaa–3′		
C12R	5′–ttcttaagctttcgccgcgccgcggattttacatatattaataaaaattaa–3′		
C13R	5′–ttcttaagctttcgccgcgccgcggcattttacatatattaataaaaattaa–3′		
C14R	5′–ttcttaagctttcgccgcgccgcggccattttacatatattaataaaaattaa–3′		
+3R	5′–ttcttaagctttcgccgcgccgcggctagattttacatatattaataaaaattaa–3′		
+5R	5′–ttcttaagctttcgccgcgccgcggctagcaattttacatatattaataaaaattaa–3′		
+7R	5′–ttcttaagctttcgccgcgccgcggctagcagaattttacatatattaataaaaattaa–3′		
+9R	5′–ttcttaagctttcgccgcgccgcggctagcagacgattttacatatattaataaaaattaa–3′		
+11R	5′–ttcttaagctttcgccgcgccgcggctagcagacgttattttacatatattaataaaaattaa–3′		
RepScreen	5′–aagctttcgccgcgccg–3′	747–731	

2.23.5 Oligonucleotide used to amplify the promoter region:

Opc16	5′–gtagtcggatatggtaacat–3′	803-784	b
Opc-Pro	5′–cattcccataaaaaacagaaaaccaagtgagaa–3′	638-670	b

2.23.6 Oligonucleotides used for mutagenesis of the IHF consensus sequence:

IHF-F	5′–aagtgagaatcgcgatcggttgtaaacaaataactatt–3′	662–699
IHF-R	5′–tgtttacaaccgatcgcgattctcacttggtttcgt–3′	689–657

Oligonucleotides marked with a 'b' indicate primers that were also obtained biotinylated. In the case of Opc12/20 and Opc19/21 the lower and higher numbers indicates the names of the un-modified and biotinylated oligonucleotides respectively.

The approximate positions of the oligonucleotides in the sequence of the *opc* region (as shown in figure 7.1) are illustrated in figure 2.3.

Figure 2.3 Showing the position of oligonucleotides in the sequence shown in figure 7.1.

Chapter 3

Mathematical models of phase variation

NOTE ON CONTRIBUTION

The work described in this chapter is part of an ongoing collaboration with Dr Mike Gravenor who is a mathematician. I recognised that there was a problem leading to the systematic over-estimation of phase variation rates by the method being used to determine mutation rates experimentally. I investigated this in the literature and identified that the problems were related to the high rates associated with phase variation, the non-linear nature of the process due to back mutation, and the impact of fitness differences. I was able to express the changes in the population algebraically but did not have the mathematics skills to convert this into a practically useful formula for the calculation of mutation rates. I could not incorporate the effects of fitness differences. At this point I approached Mike. He was able to derive the formulae that are described and used in this chapter. His questions when designing these formulae raised the issue of discontinuous and continuous models of change. Consideration of this question led me to appreciate that the type of mutational process that mediates phase variation determines which of these models is the most appropriate. In summary, Mike contributed all of the derivation of the mathematical models described in this chapter. The recognition and definition of the problem, the application of the models to phase variation, the simulations, and their interpretation are entirely my own work.

3.1 The determination of mutation rates associated with phase variation

3.1.1 Phase variation rates and bacterial fitness

Phase variation describes a process of reversible, high frequency phenotypic switching that is mediated by DNA mutations, re-organisation or modification. The rate of phase variation will affect the population in a number of ways. It has been proposed that there is an optimal rate of phase variation that occurs as the reciprocal of the frequency of transition between the alternate environments for which the varied phenotype is adaptive (Moxon *et al.*, 1994). It can also be argued that during certain critical steps of transmission and colonisation, for example in a small inoculum that has to establish infection, diversification at a high rate which permits colonisation may be adaptive. The principal requirement for success for a microbial parasite is to survive as a propagating population in the host (Finlay & Falkow, 1989). The energy expended in this process or even the proportion of the population that is lost is irrelevant, as long as the residual population can expand to survive and be transmitted to new hosts (Wise, 1993). Hosts present an array of distinct niches that are subject to change over time, including adaptive immune responses. Phase variation is a means that has been adopted by several bacterial species as a mechanism by which to generate the diversity within bacterial populations that increases

bacterial fitness. Being able to determine the rate at which these processes occur and the nature of any factors that influence them is integral to understanding the impact of these processes on the evolution and dynamics of the population as a whole and the host-bacterium interaction. To do this, a means with which to reliably determine and compare phase variation rates within and between experiments and bacterial populations is needed.

3.1.2 Luria and Delbrück, Lea and Coulson, and Stocker

The estimation of mutation rates predominantly uses methods derived from a classic paper by Luria and Delbrück (1943). In these studies a number of similar cultures are grown under conditions that are as near identical as possible. The cultures are started with an inoculum of cells of the same genotype. As the cells divide some may give rise to clones of mutants. At the end of the experiment the cells are plated out and the number of cells which have the phenotype of the mutation of interest is determined. Because mutation can occur at any time during the culture the number of mutant colonies at the end of the experiment represents the number of new mutations and the accumulation of mutants from the replication of those that arose prior to the final round of division in the culture. In some cultures mutation will occur earlier than in others. The earlier this event occurs, the larger the proportion of mutants that will be present at the end of the experiment. The occurrence of these so called 'jackpot cultures' due to early mutations are an inherent feature of the stochastic nature of this process. It was this insight that led to the conclusion of Luria and Delbrück that mutations (to bacteriophage resistance) were occurring prior to exposure to the selective pressure for which they were adaptive. They showed that the variance between replica experiments was much greater than (as opposed to equal to) the mean, and the distribution of the number of mutants was characterised by a long tail of rare cases of high numbers of mutant bacteria. Their analysis demonstrated that the mutations were spontaneous and their approach was used by others to show similar spontaneous generation

of mutants in antibiotic resistance (Demerec, 1945) and ultraviolet radiation survival (Witkin, 1946).

Based upon the study of the distribution of mutants in their cultures, Luria and Delbrück described two methods to determine the mutation rate (μ). One used the percentage of cultures in which no mutants could be detected (P_0) and the other used the mean number of mutants in the final culture (m). The P_0 estimator uses the observation that if the probability of a given mutation occurring in a culture is μ then the number of new mutations in a culture will be well approximated by a Poisson distribution with the parameter $m = \mu N$, where N is the final number of cells in the culture. Accordingly a proportion of $P_0 = e^{-m}$ cultures will give rise to no mutant colonies. It follows that a number of new mutants, m_0, can be derived from P_0, using $m_0 = \ln (C/z)$, where of C cultures, z cultures are devoid of mutants. The method, based upon the average number of mutants in a final culture uses the equation $\mu = (M/N) / g$. This determines the mutation rate from the increase in the number of mutants (M) compared to non-mutants (N) in the final population, and the number of generations in the experiment (g), assuming that this is a linear function. Their insight into the nature of the distribution also implied that, at least theoretically, the mean is a poor indicator of m from which to determine μ. Despite this the method that they presented using the mean has been observed in practice to be not as poor as it is often described (Stewart, 1994). In addition, the coefficient of variation of m is also the coefficient of variation of μ. In other words if we know m to within 10% we also know μ to within 10% (Jones *et al.*, 1994).

Luria and Delbrück developed a qualitative description of the distribution observed in fluctuation tests and did not derive the shape of the distribution to be expected on the basis of the spontaneous mutation theory. This may have been because it was unnecessary for their purposes, but may also have reflected an appreciation that the necessary assumptions may not accurately reflect experimental conditions and that divergence from such a

distribution would not affect the interpretation of their results. Lea and Coulson (1949) sought to extend the model of Luria and Delbrück by calculating the form of the distribution of numbers of mutants in parallel cultures to be expected from the spontaneous mutation theory. Their analysis also facilitated a consideration of the methods used to calculate mutation rates. Their most important contribution in this context is the recognition that the right-handed skew seen in these experiments confirms that the mean number of mutations is a relatively poor basis for calculation of the mutation rate and that the median value is more reliable. They also empirically determined an equation to generate a more accurate value for use in these experiments called the 'median estimator'. They observed that the probability, P_r, of their being r or fewer new mutants in a culture in which the expected number of new mutants is m, is well approximated, when r >> 1 by a function of $[r/m - \ln(m)]$. They found empirically that P_r assumes the value of 0.5 when $[r/m - \ln(m)] = 1.24$ and thus that a number of new mutants, m_{med}, can be derived from the observed median number of mutants, r_m, when $r_m/m_{med} - \ln(m_{med}) = 1.24$. Lea and Coulson also stated that the mean of a large sample will give no better estimate of the mutation rate than a single observation. This contention is not supported by simulation experiments (Stewart, 1994). Whilst doubling the sample size does not cut the standard deviation by a normal proportion of $\sqrt{2}$, increasing the number of mutants that are included in the results by 4 fold does reduce the standard deviation by a factor of 1.25. Lea and Coulson's solutions were computationally complex, as are many of the subsequently described alternatives, and the interpretation of the results requires the use of probability tables to determine the most likely mutation rates.

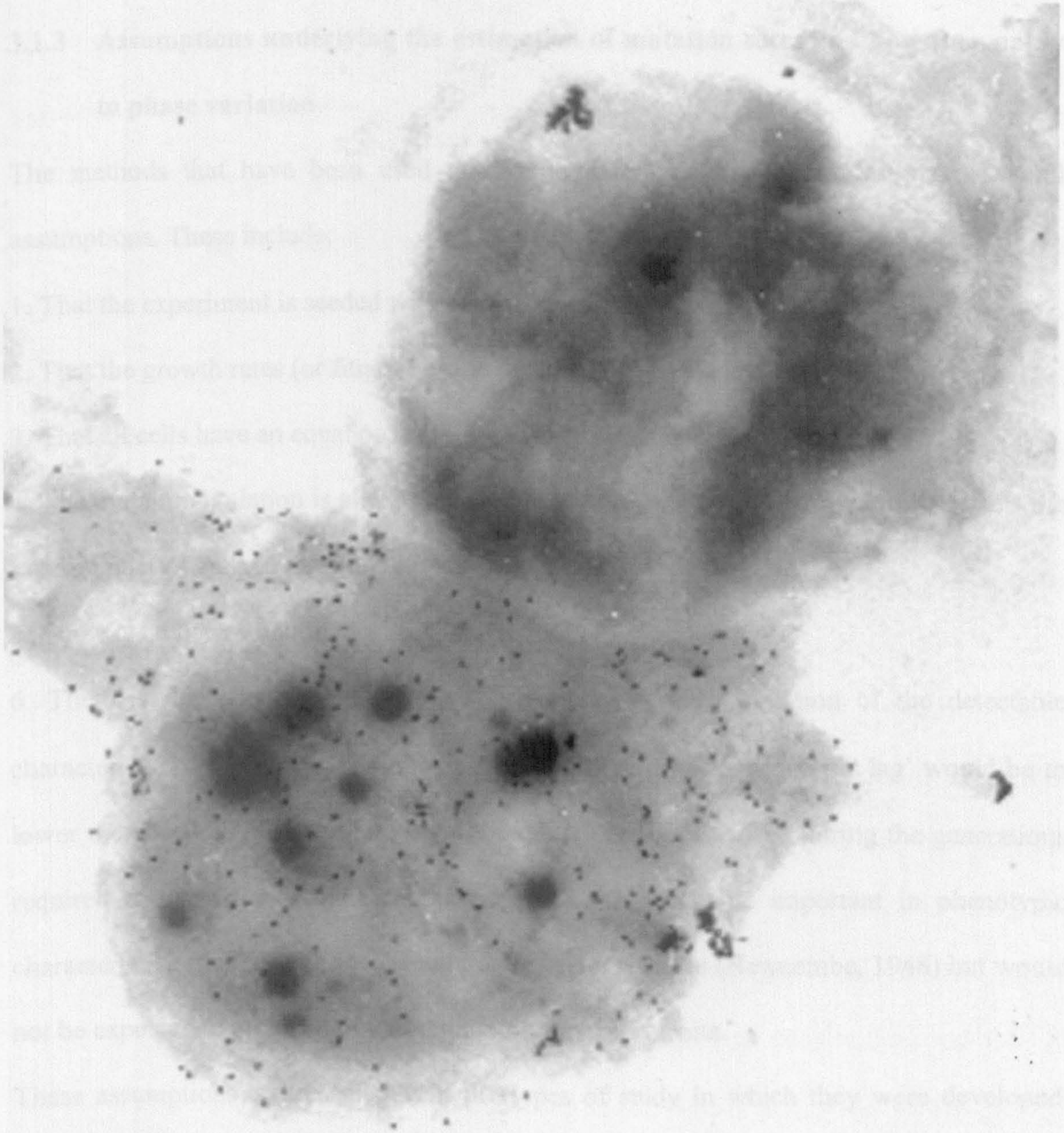
The first paper to specifically consider rates of phase variation was a study of *Salmonella* flagellae by Stocker (1949). His detailed description of the method of Luria and Delbrück reveals a correction necessary when considering mutations in which both progeny do not have the mutant phenotype being studied and this highlights essential differences between the mechanisms of phase variation. The mutation rate formula $\mu = (M/N) / g$ is based upon

a model in which after one generation from a population of n cells there will be $2n$ total cells and $2\mu n$ mutants. This is appropriate for the model system that he was studying. In this instance a DNA re-organisation occurs prior to division which is subsequently present in both progeny. However, slippage-like processes in repetitive DNA sequences are the most common mechanism that mediates phase variation. In this case variation is most likely to occur at the time of DNA replication, in which instance it is unlikely that a mutation will occur independently on both replicated chromosomes. An example of phase variation that illustrates this is shown in figure 3.1. Consequently only one of the progeny cells will generate the alternate phenotype so the real number of mutants would only be μn after one generation. This means that in the study of this type of mutational process the expected proportion of cells after g generations is not $g\mu$ as stated, but $g\mu/2$, leading to a two-fold underestimation of the mutation rate. In addition, the inversion that mediates fimbrial phase variation can occur at any time in the life cycle of the bacterium, which contrasts with the slippage mediated events that are most likely to occur at division. Stocker investigated the concordance between the observed increase in mutants and the linear accumulation predicted by the $\mu = (M/N) / g$ model. Whilst the results were interpreted as consistent with a linear increase they do not exclude alternative interpretation, especially in the case studied in which mutations occurred to restore a mutant phenotype to a wild type. As the number of generations in an experiment increases the number of such back-mutations increases. In experiments studying very low mutation rates in processes that are not inherently reversible this introduces an insignificant error. Stocker recognised and confirmed experimentally, as previously stated by Bunting (1940), that as the number of generations increases the population approaches an equilibrium determined by the ratio of forward and backward mutation rates in an exponential fashion. This effect increases with the rate of mutation. Once mutation rates in both directions and the equilibrium conditions were known he was able to use this to make a rough correction for the non-linear increase in the variant population. Although not quantified, Stocker's

Figure 3.1

Phase variation of *Neisseria meningitidis* Opc protein observed by immunogold electron microscopy. Individual cells of a diplococcus can be seen with phase on (presence of electron dense particles on surface) or off, demonstrating that only one of a pair of dividing cells has the variant phenotype.

Photograph kindly provided by Professor M Virji and Dr DJP Ferguson.



observations were also amongst the first to emphasise the importance of the influence of differences in the growth rates of the mutants upon the final proportion of mutants in a culture, and hence on the determined mutation rate. He was also the first to encounter the problems of having to initiate cultures with single cells associated with systems with high mutation rates, including increasing the number of generations in the experiment that may generate 'jackpot cultures'.

3.1.3 Assumptions underlying the estimation of mutation rates and how they apply to phase variation

The methods that have been used to date to determine mutation rates make several assumptions. These include:

1. That the experiment is seeded with cells of a single phenotype / genotype.

2. That the growth rates (or fitness) of the different genotypes are equal.

3. That all cells have an equal probability of mutating (including the mutants).

4. The mutant population is always small and the total number of wildtype cells equals the total number of cells in the culture.

5. There is no back-mutation.

6. There is no 'phenotypic lag'. This is a delay in the expression of the detectable characteristic of the mutant (Armitage, 1952). The effect of 'phenotypic lag' would be to lower the observed mutation rate since the mutations that occurred during the generations required for detection would not be observed. This may be important in phenotypic characteristics such as bacteriophage or antibiotic resistance (Newcombe, 1948) but would not be expected to affect colony immunoblotting experiments.

These assumptions are reasonable in the types of study in which they were developed. However, the high-frequency, reversible nature of phase variation provides a more challenging context in which to perform this type of study. The experimental approach used by Stocker, in which serial liquid cultures in broth were used had the advantage of

being able to include a large number of generations over a period of days in order to increase the proportion of mutants in the population. Large inocula were used in serial subculture to ensure that the proportion of mutant cells in the inoculum was representative of the parent culture. However, in many cases in order to ensure that an experiment is initiated with cells of a single phenotype, this frequently has to be a single cell. There is often an additional constraint in that the number of divisions may be limited to those that can be generated upon solid media. This increases the number of generations prior to the population achieving a size of $1 / \mu$ and hence the number of divisions prior to the population achieving a size in which a mutation is likely to occur. The number of mutants will remain at zero for a greater number of divisions than when a larger inoculum is used and hence the final proportion of mutants will be reduced. This latter problem was addressed by the correction of Drake (1991) (see below) but the problem of a reduced number of detectable mutants remains. In addition, early studies used methods of detection that permitted the screening of whole cultures for the mutant phenotype of interest. When other detection methods, such as immunoblotting, have to be used this is usually not possible, which reduces the sample size of the population and the number of detectable mutants.

The two most important remaining factors are effects of differences in fitness between variants being studied and the influence of back-mutation. These are of particular concern in the study of the phase variation of contingency genes. By definition, contingency genes confer potential fitness changes with respect to the different environmental conditions encountered by the population of bacterial cells. It is not reasonable to assume that variants are phenotypically equivalent and it is also necessary take into account the conditions for which they are adaptive. Finally, as observed by both Bunting (1940) and Stocker (1949), back-mutation can have a significant contribution in phase variation experiments.

3.1.4 A brief review of currently used methods for phase variation rate determination

Despite this extensive background work, the study of phase variation has not consistently made use of the methods that have been developed. The use of different methods and a lack of appreciation of their underlying assumptions has led to confusion in the use of terms and a situation in which the results of similar experiments cannot be compared. This has made it impossible to compare rates generated by different investigators and to compare rates between different phase variable systems. This is compounded when the method used or the primary data from which the rates were determined are not explicitly stated. For example, Foster (1999) used a fluctuation test but did not reference the method or state whether m or P_0 was used to calculate the mutation rate. A lack of attention to detail in this area has led to an abundance of erroneously high descriptions of phase variation rates in the literature which hamper the understanding and investigation of this important mechanism of adaptive bacterial diversification.

The study by Stocker (1949) laid an excellent foundation for the investigation of mutation rates in phase variation. According to their descriptions other early reports of phase variation rates used the $\mu = (M/N) / g$ equation in the study of pilus variation in *E. coli* (e.g. Eisenstein, 1981) which they used on the basis of Enomoto and Stocker (1975) which was in turn based upon Stocker (1949), and Luria and Delbrück (1943). However, Eisenstein (1981) did not state how many replicate cultures (if any) were used in their studies, and it seems likely (he does not state explicitly) that the mean values were used in their experiments. Nor do they do so in subsequent papers (Eisenstein *et al.*, 1987).

Some papers avoid the issue of calculating the mutation rate, μ, and simply describe the number of mutants. For example, in a recent study of the influence of promoter strength upon phase variation rates, the percentage of variants in a population was determined by suspending single colonies, plating to obtain approximately 500 colonies, and counting the

number of variants (Belland *et al.*, 1997). There are problems with this approach. Whilst the proteins studied form surface pores and phenotypic changes involve the gain or loss of these pores, it is assumed that this has no effect upon growth rate. The number of cells in the suspended colonies, and hence the number of generations was not assessed. Since the proportion of mutants in the culture increases with the number of generations, this is a serious potential confounding influence when there is less than three-fold difference between the phenotypes deemed least and most mutable. Secondly, this data does not provide any means by which the data can be compared with the results of any other study of either the same or different phase variable genes.

Several studies fail to distinguish between m and μ. For example, Roche *et al.*, (1994) in a study of LPS phase variation in *Haemophilus influenzae* grew cultures to log phase (number of generations undetermined) and then sub-cultured a proportion of the cells to obtain single colonies for immunoblotting. These figures (a value of m) were erroneously described as a rate of approximately 1% of bacteria per generation (i.e. as a value of μ). This same method has even been used in cases where the phenotypes are stated to have different growth rates (Weiser, 1993). This error is widespread and is an important source of error in the description of phase variation rates at inaccurately high rates. Other papers, even ones cited for their methods of rates determination (e.g. Hammerschmidt *et al.*, 1996a, cited by Bucci *et al.*, 1999), are not explicit about essential steps in the investigation of mutation rates. These papers serve to illustrate common problems with this type of study. Even though these authors state that mutation rates are expressed as the frequency of detectable mutations in a culture of 10^4 or 10^5 cells, this is not an appropriate use of the term 'mutation rate', unless it is related to the appropriate number of generations. An approach that does not differentiate between m and μ and describes the result as a rate is misleading and suggests that the rate is higher than it actually is. In addition, not only are the number of divisions being assessed not quantified in these studies, it is also not explicitly stated whether the cultures being investigated have been

subjected to subculture or if the experiments are performed starting with single colonies. Finally, the number of repeated cultures are not stated. 'Rates' determined in this way have even been directly compared with rates determined on a per cell per generation basis (method of determination of *m* used to determine μ unstated) (Bucci *et al.*, 1999) – an approach which will falsely generate or exaggerate differences between the two processes of mutation being addressed.

3.1.5 Practical requirements for a method to determine phase variation rates and the two-step nature of the process.

Interest in this field increased following the studies of Cairns *et al.* (1988) in which evidence was presented showing that the observed distribution differed from those predicted by Luria and Delbrück under certain conditions. These results were interpreted to suggest that bacteria have mechanisms for generating mutations that are 'directed' by the selective conditions to which they are exposed. This led to reconsideration of the methods and various ways of improving the description and analysis of the distribution of the mutants generated in these experiments (e.g. Stewart *et al.*, 1990). However, these analyses were not focussed upon the specific issue of determination of mutation rates and if this is seen as a separate issue then it is not clear that it should be pursued by comparing distributions in this fashion.

The experimentalists in this field are unlikely to utilise mathematically complex solutions to this problem. They do require a practical method that yields a reasonable approximation to the actual mutation rate, μ. In addition to the practical experimental constraints that apply to the mutational process and phenotypic changes that are under investigation, there are two stages to the process that have to be considered separately. The first is an analysis of the experimental data to generate a value (m) from which the mutation rate (μ) can be derived. The second is a calculation of μ from this data. The first stage is the one that has received most attention to date. Once this two-stage nature is appreciated it becomes

possible to derive new methods that, although generally applicable, allow consideration of factors important to quantitating phase variation. The major confounding factors are differences in bacterial fitness and back mutation.

In this context, methods addressing the distribution of mutant accumulation in parallel cultures have been developed to study phenotypes with differing fitness. The first example was that of Koch (1982), a complex and computationally intense solution. Subsequently, Stewart *et al.* (1990) also introduced formulae for the prediction of mutant distributions that could be used when the growth rate of the mutant and the wild-type differ. However these methods, which account for the influence of 'biological parameters' do so by modifying the equations that determine the Poisson parameters and do not lead easily to methods that will help in the practical determination of mutation rates under these circumstances. However, these studies are interesting because they highlight the fact that when the assumptions inherent to the Luria and Delbrück method are not concordant with the experimental conditions (which is often the case) then divergence from their predicted distribution is to be expected, and does not necessarily indicate the action of 'directed' mutation (Mittler & Lenski, 1992). Subsequent work has resolved many issues related to the prediction of the expected Luria and Delbrück distribution using predetermined mutation rates (Sarker *et al.*, 1992). However, this also does not provide a computationally simple solution to the determination of mutation rates from experimental data.

3.1.6 The determination of P_0 or *m*

Although it is one of the most efficient computationally easy estimators of the mutation rate (Li and Chu, 1987), one of the two Luria and Delbrück alternatives for determining the rate of mutation, P_0, cannot be used in most phase variation experiments. This method assumes that all of the cells in a culture are plated and all mutants can be detected. When a non-lethal selection/detection method is used, such as colony immunoblotting, this investigation of a whole culture is frequently impracticable. It also requires many iterations

of the experiment which may also not be possible. For example, it has been estimated that for an optimally designed Luria and Delbrück type analysis with a 100% plating efficiency from 43 to 39 cultures are required depending upon the method used (Kendal & Frost 1988; Jones *et al.*, 1994). However, when a constraint of only being able to selectively plate 10^6 cells from a culture of 10^7 was considered, it expands the number of selective plates to 430, 380 or 76 plates depending upon the method used. This constraint is small when compared to the ability to enumerate and detect variants on plates that only allow accurate data collection from 10^4 cells derived from cultures with greater than 10^8 cells which is a typical situation in the study of bacterial populations. Most importantly, the use of P_0 requires that a significant proportion of cultures do not contain mutants, making this method unsuited to the investigation of phenomena in which the mutation rate is high. A situation typified by studies of phase variation. An alternative estimator has been derived when few or none of the final cultures has 0 mutants in which plating of only part of each culture is used to generate mutant free plates. The amount of dilution necessary to achieve this is used in the determination of the average number of mutants in the cultures (Jones *et al.*, 1994). This solution is more complex and the experiments have to be titrated so that no mutants are present in a significant number of cultures. It also adds an additional source of variability due to sampling because of the stochastic nature of the events being studied.

The methods that have been used to estimate m are: the mean, the median, the 'median estimator', maximum likelihood, and Bayesian modelling. A comparison of the different methods for generating the value of m in mammalian cell culture experiments has indicated that when there is sufficient data, the performance of the methods decreases from maximum likelihood, the P_0 method, the median method (using the true median), the upper quartile method, to the mean method (Li & Chu, 1987). The authors conclude that the mean and upper quartile methods should be abandoned (the latter method was designed to compensate for phenotypic delay (Armitage, 1952 & 1953) but is the most susceptible to 'jackpot cultures') in the presence of the more accurate methods. The estimated value of μ

was found to vary up to 3 fold between the methods and the accuracy increased with the number of generations included in the experiment. However it should be noted that the largest difference between the best method and the median method was only 28%. They conclude that of the computationally easy estimators the median is the most efficient determinant of *m*.

In a subsequent study of a large number of simulated experiments, the estimators were compared using predetermined mutation rates, as were factors that affected the accuracy of the rate determination (Stewart, 1994). Estimates of the rates of variation improved with both increasing sample size and with the number of mutants that are observed in each experiment. The majority of simulations gave an approximately correct result. The simulation experiments support the maximum likelihood measurement as the best predictor of mutation rate. The other methods, particularly the median estimator of Lea and Coulson, performed acceptably well and did not normally introduce errors of greater than 10% (and frequently much less). However, a few of the estimators gave a mutation rate an order of magnitude too high due to the inclusion of 'jackpot cultures'. This is easily within acceptable limits for the study of many aspects of mutation and was in a context of inclusion of all results from the simulated cultures in each experiment, i.e. including all of the 'jackpots'. The maximum likelihood analysis can be considered to be a special case of Bayesian estimation technique and Bayesian procedures have now been presented to analyse fluctuation studies, but these are computationally intense. Each of these methods still requires a relatively large dataset. The Bayesian models were used to investigate the effect of 'jackpot' cut-offs. This showed that a cut off of 3 to 4 fold the median number of mutants did not significantly affect the accuracy of the estimates and the lower the mutation rate (number of detected mutant colonies) the lower the acceptable cut-off (Asteris & Sarker, 1996). However, the increase in the median due to 'jackpots' is not addressed by counting them as a reduced value.

3.1.7 The exclusion of results from jackpot cultures

Luria and Delbrück stated: 'there is no general criterion by which one might eliminate such cultures (jackpots) from the statistical analysis, because, in a culture with an exceptionally high count of resistant bacteria, these do not necessarily stem from one exceptionally early mutation, but may also be due to an exceptionally large number of mutations after (the start of the experiment)'. This view was appropriate at the time because little was known of the processes being studied and because they were primarily interest in the distribution of the mutants in different cultures. However, the typical purpose of the mutation rate experiments is not to study the distribution of variance within the replicate cultures that are used, it is to determine a sufficiently reliable mutation rate for the contribution of the variable phenotype under investigation to population diversity and fitness to be assessed. An awareness of the practicalities of performing these experiments and an understanding of the nature of the process being studied should permit rational and sufficiently accurate determinations of mutation rates. The purpose of the experiments is to determine the rate of mutation within a cell population. This can be determined from the measurable accumulation of variants in populations that are sufficiently large and do not initially contain mutants. When an initial starting culture is not sufficiently large (i.e. less than the reciprocal of the mutation rate: $1 / \mu$) it is prone to the generation of jackpot cultures. When an inoculum of less than $1 / \mu$ has to be used then the number of generations **and** the mutants in the final population that are descended from events that occur during these preliminary divisions essentially lie outside of the experiment. This was recognised by Drake (1991) when he adjusted the number of generations used to determine μ from m (see below). By the same argument it also follows that the 'jackpot cultures' are the product of events that are outside of the experiment and should also be excluded. If this is accepted then the distribution will approach Poisson and the practical and statistical objections to the use of the mean (or median or other value of m) cease to be problematic.

In the majority of experiments the issue is one of seeking the normal mutation rate in the population being studied. This is the mutation rate in the population that are not hypermutators and in populations in which the number of mutants is representative of mutation within the population and not altered by the founder effects of an early mutation. It is therefore reasonable to exclude results generated by phenomena that are clearly understandable from a biological perspective and which are not representative of the processes being studied. The right-handed skew in the distribution of mutants seen in these experiments has been recognised from the very beginning of this type of study and it is an inherent property of the stochastic process being studied. As an alternative to struggling with complex mathematical solutions it is pragmatic to exclude the unrepresentative data. Using a median value determined from a small series of cultures and excluding values of 3 times this value could do this, for example. If there is a clear clustering of results then since it is extremely improbable that these represent multiple 'jackpot cultures', there is no need to collect substantially larger data sets for every experiment. Experiments of this type are frequently conducted in series and the normal range of the results of these experiments is likely to be known. Multiple 'jackpot' cultures are unlikely and when results lie within the expected range then 3 to 6 consistent results are probably sufficient to determine the typical mutation rates, and to recognise and exclude jackpots.

3.1.8 The second step

The only attempt to date to address the second stage of the determination of mutation rates is that of Drake (1991). In the experimental situation it is assumed that cultures are initiated from small mutant-free inocula and are grown extensively, and their mutant frequencies are then determined. In this situation it has been argued that N_1 is not the initial inoculum but rather the value when the population reaches the size when a mutation is likely to occur (i.e. $1 / \mu$). In this special case in which $f_1 = 0$, then $\mu = f / \ln(N\mu)$. The formula $\mu = (f_2 - f_1) / \ln(N_2 - N_1)$ is used where f = the mutant frequency (at time 1 or 2)

and N = the population size (at time 1 or 2) instead of $\mu = (M/N) / g$. This correction, taking account for the replications in the early period of the experiment is a sensible refinement and yields higher mutation rates than if these replications are included. The equation must be solved by iteration. It is also not clear why in this study the median value (used to avoid the sensitivity of the mean to 'jackpot' cultures) is used but in such a way that all data, including jackpots is included. Having specifically excluded the divisions that generate 'jackpot' cultures it is unreasonable to include them in the final data used to determine the mutation rate.

Both the original equation and that of Drake have similar assumptions: 1. that the increase in the mutants in the population is linear, 2. that the number of mutants in the population is negligible, 3. that there is no significant contribution of back-mutation, 4. that the fitness of the two phenotypes is equal. These assumptions are not applicable to the analysis of phase variation in which the mutations occur at a relatively high-frequency and are also reversible. They do not allow studies to be conducted under the conditions of selection for which the alternate phenotypes are adaptive. They are also unnecessary and can be accounted for in the second-step calculation.

3.1.9 A new discrete model for phase variation (contributed by Mike Gravenor)

$$A_{n+1} = 2(1-\rho)A_n + \rho A_n + \sigma B_n$$
$$B_{n+1} = 2(1-\sigma)B_n + \sigma B_n + \rho A_n \qquad \text{Model 1}$$

This model describes synchronous growth of a population of bacteria. A_n and B_n represent the numbers of original phenotype and mutant phenotype at generation n. Switching can occur from $A \rightarrow B$ and $B \rightarrow A$, potentially at different rates. The probability of a bacterium of type B arising from division of a type A bacterium is ρ, hence the discrete switching rate is ρ *per* generation. The 'back-switching rate' is σ *per* generation. At division, if switching from A *does not* occur (probability 1- ρ) 2 progeny of type A arise. If a phase variation occurs (probability ρ), the progeny consist of one type A and one type B.

The model can be extended to include fitness differences between the two types, expressed as the probability of surviving to division. Here, proportions d_A and d_B of types A and B respectively survive each generation.

$$A_{n+1} = 2d_A(1-\rho)A_n + d_A\rho A_n + d_B\sigma B_n$$
$$B_{n+1} = 2d_B(1-\sigma)B_n + d_B\sigma B_n + d_A\rho A_n$$

Model 2

Starting from an initial number of normal and mutant types (A_0, B_0), these models can easily be iterated on a spreadsheet in order to determine the expected proportion of normal and mutant phenotype after any number of generations. This value will be governed by the parameters (switching rates and fitness differences) and the initial proportions of A and B.

The solutions for A and B after n generations can be also be obtained directly. With no fitness differences,

$$A_n = \frac{1}{\rho+\sigma}\left(2^n\sigma(A_0+B_0) + (2-\rho-\sigma)^n(\rho A_0 - \sigma B_0)\right)$$

$$B_n = \frac{1}{\rho+\sigma}\left(2^n\rho(A_0+B_0) + (2-\rho-\sigma)^n(-\rho A_0 + \sigma B_0)\right).$$

Starting with a culture consisting solely of type A, the proportion of mutants after n generations is

$$m_n = \frac{\frac{\rho}{\rho+\sigma}\left(2^n - (2-\rho-\sigma)^n\right)}{2^n}.$$

From an observation (or summary observation such as the median) of m_n, both switching rates cannot be calculated. Hence an assumption about their relative values must be made.

Assuming no back switching ($\sigma = 0$)

$$m_n = 1 - \left(1 - \frac{\rho}{2}\right)^n,$$

hence the switching rate can be estimated by

$$\rho = 2\left[1 - \sqrt[n]{1 - m_n}\right].$$

Assuming equal forward and backward switching rates ($\rho = \sigma$)

$$m_n = \frac{1}{2} - \frac{(1-\rho)^n}{2},$$

hence the switching rate can be estimated as

$$\rho = 1 - \sqrt[n]{1 - 2m_n}\,.$$

If differences in fitness are included in the model an analytical expression for m_n can also be derived. Due to the large number of terms, this calculation is best done with a spreadsheet.

Using the same notation the Luria and Delbrück model is essentially

$$A_{n+1} = 2A_n$$

$$B_{n+1} = 2B_n + 2\rho A_n$$

(i.e. no back switching and the population size A is approximately the total population size n).

Aside from the fact that one mutation is assumed always to lead 2 new mutant progeny, the use of *m/n* as an estimate of the mutation rate is very good, particularly at low mutation rates, and even if back switching occurs. The inclusion of fitness differences (and high mutation rates) can however skew the results considerably and it is best to use model 2 above. All the parameters in this type of experiment can't easily be estimated. It is important to state the assumptions of the method used. However, if different systems are likely to vary in these unknown parameters, it is very difficult to meaningfully compare rate of phase variation. The clear statement of the assumptions used in these calculations will at least make this more apparent.

3.1.10 Conclusions

The study of phase variation rates needs standardisation. It is essential to distinguish between the number of mutants present in the culture and the number of mutations that occurred in the generation of the culture population when describing mutation rates and values of *m* must not be described as mutation rates (μ). Giving values of *m* as mutation rates is misleading, leads to misleading impressions of high phenotypic variability, and is impossible to rationalise with the selection required to maintain these traits in those bacteria that have multiple phase variable characteristics. Consequently, all experiments that study rates must include a determination of the number of generations studied. Wherever possible they should also include an assessment of the relative fitness of the phenotypes being investigated. The median is an acceptible determinant for *m* for most purposes and results from 'jackpot cultures' may be reasonably excluded. If greater accuracy is necessary then theory is available and the Bayesian model is the best alternative. The proposed model is designed to force investigators to be more explicit about the assumptions underlying their rate calculations. Given a relative fitness difference and assumed relative rates of mutation for the populations under study, a simple formula expresses an estimate for the per generation probability of switching from a proportion of mutants in the final culture. It also facilitates the study of phase variation under selective conditions once either the mutation rate has been determined under non-selective conditions or when the relative fitness of the phenotypes in the experimental conditions is known. Whilst this formula allows for the investigation of mutation rates in phase variation it is not specific to this type of mutational study and can also be applied to mutation rate experiments generally.

3.2 The influence of phase variation and selection on population structure

3.2.1 Introduction

The relative contributions of the rate of phase variation and associated alterations in bacterial fitness to population structure have not been previously studied. The model developed in section 3.1, since it allows for both fitness differences and the back-mutation typical of phase variation, facilitates the modelling of this process.

Papers which have determined phase variation rates in *H. influenzae* using appropriate methods have described mutation rates of 10^{-5} to 10^{-4} (De Bolle *et al.*, 1999). Analysis of the phase variation rate of *opc* in *N. meningitidis* using the new model also reveals a similar rate of phase variation of 1.2×10^{-4} . These results are consistent with re-analysis of data from other studies of phase variation using the model described in section 3.1. There are frequent descriptions of higher rates associated with phase variation in the literature. In some instances the origin of these rates are not identified (e.g. Weiser & Pan, 1998) and others cite papers that in turn cite others that do not include a rate determination (e.g. Weiser *et al.*, 1998a). Those papers that have reported rates in the region of 10^{-2} have not actually been determining rates per generation per cell (μ) but have instead based their observations upon the accumulation of variants in cultures over multiple generations (m) (typically the number of generations from single cell to bacterial colony). As discussed in section 3.1, this will consistently over-estimate what is described as a rate of variation and this probably underlies most if not all of the high rate descriptions of phase variation. For this reason, rates that are truly reflective of those normally involved in phase variation in the range of 10^{-3} to 10^{-5} have been used in the simulations described in this thesis.

3.2.2 A continuous model of phase variation (model contributed by Mike Gravenor)

As described in section 3.1.2, there are differences in the mutation processes that lead to phase variation depending upon whether the events occur primarily during cell division (resulting in a single variant progeny) or during the whole of the cell cycle (typically

resulting in two variant progeny). These processes are best described by discrete and continuous switching rates respectively. The model described in section 3.1 covers the discrete switching situation. Here a model to describe a continuous switching process is introduced. The purpose of this section is not to describe a specific phase variable system, but to describe the impact of phase variation upon population composition and the behaviour of phase variable populations in models with different switching rates and fitness differences. The general conclusions of the simulations described below apply equally to the continuous and the previously described discrete model. However, if rates are to be estimated from particular datasets it is important that the most suitable model is selected for this purpose. For a given dataset, the rates derived using these two models will not be the same. This model can also be run in a spreadsheet such that values for the relative fitness and mutation rates can be varied. The equations are as follows:

$$\frac{dX_t}{dt} = -\beta X_t + \alpha Y_t + R_x X_t$$

$$\frac{dY_t}{dt} = \beta X_t - \alpha Y_t + R_y Y_t$$

X_t is the proportion of phenotype X at time t, Y_t is the proportion of mutant phenotype. Phenotype X grows at an average continuous rate R_x, Y at the rate R_y. For each population to double over a generation time t, $R_x = R_y = ln(2)$. The continuous per capita switching rate from $X \rightarrow Y$ is β, the average switching back rate from other phenotypes is α. These rates are not probabilities, but are such that the per capita probability of a switch occurring (e.g. $X \rightarrow Y$) during the very small time interval Δt is $\beta^* \Delta t$.

The equations can be solved to give an expression for $Y/X+Y$ (i.e. m) or the rates, at any time t. The general solution for the full model is.

$$X_t = v_1 k_1 e^{r_1 t} + v_3 k_2 e^{r_2 t}$$

$$Y_t = v_2 k_1 e^{r_1 t} + v_4 k_2 e^{r_2 t}$$

where the following are the eigenvalues of the matrix of parameters:

$$2\gamma_1 = (R_1 + R_2) - (\alpha + \beta) + \sqrt{(R_1 - R_2)^2 + (\alpha + \beta)^2 + 2(\alpha - \beta)(R_1 - R_2)}$$

$$2\gamma_2 = (R_1 + R_2) - (\alpha + \beta) - \sqrt{(R_1 - R_2)^2 + (\alpha + \beta)^2 + 2(\alpha - \beta)(R_1 - R_2)}$$

and

$$\begin{bmatrix} v_1 & v_3 \\ v_2 & v_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \alpha^{-1}(\gamma_1 + \beta - R_1) & \alpha^{-1}(\gamma_2 + \beta - R_1) \end{bmatrix}$$

The analytical solution is also dependent on the initial proportion of the population that is phenotype X, which gives the constants

$$k_1 = \frac{v_3 Y_{(t=0)} - v_4 X_{(t=0)}}{v_2 v_3 - v_1 v_4}$$

$$k_2 = \frac{X_{(t=0)} - v_1 k_1}{v_3}$$

The above expressions can be simplified for particular assumptions. For equal growth rates ($R_x = R_y$) and no back switching ($\alpha = 0$), the proportion of type Y after time $t = m_t$ where

$$m_t = \frac{1 - e^{-\beta t}}{e^{-\beta t}}.$$

The switching rate can therefore be estimated from m_t according to the formula

$$\beta = -\frac{\ln\left[\frac{1}{1+m_t}\right]}{t}.$$

For equal growth rates ($R_x = R_y$) forward and back switching occurs, but at the same rate ($\alpha = \beta$)

The proportion of type Y after time $t = m_t$ where

$$m_t = \frac{1-e^{-2\beta t}}{2}$$

The switching rate can therefore be estimated from m_t according to the formula

$$\beta = -\frac{\ln(1-2m_t)}{2t}$$

3.2.3 Simulations and discussion

When the phenotypes have equal fitness, phase varying cultures will approach an equilibrium proportionate to their mutation rates (Bunting, 1940; Stocker, 1949). This is modelled in figure 3.2a using a mutation rate of $\alpha = ß = 0.01$. In some instances phase variation is mediated by repeats located within open reading frames such that alterations in the repeat tract length affects expression by moving the 3′ reading frame in or out of frame with the 5′ initiation codon. In this situation, if the rate of repeat variation in each direction is equal, then the observed phenotypic variation would be expected to be modelled by $\alpha = 2ß$. This is modelled in figure 3.2b, in which $\alpha = 0.01$ and $ß = 0.005$). In this situation the population approaches a 1/3 : 2/3 composition. The rate at which the equilibrium is approached is proportionate to the mutation rate. The population structure will change over time in the absence of selection, simply as a consequence of the mutational process. As this simulation demonstrates, if the mutation rates were as high as 10^{-2} then this would rapidly lead to the presence of a large proportion of variants. This does not make sense in the

Figure 3.2a

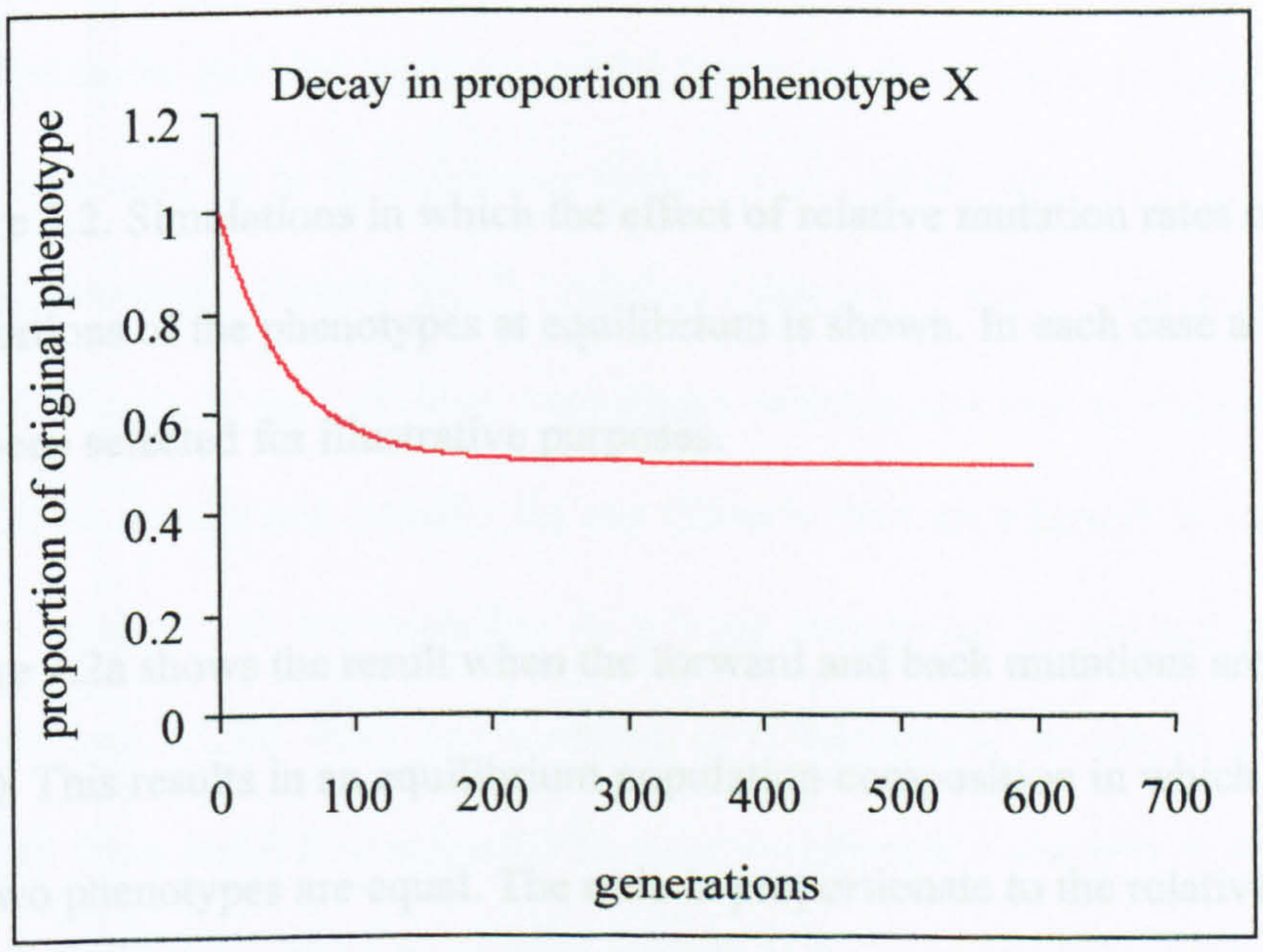


Figure 3.2b

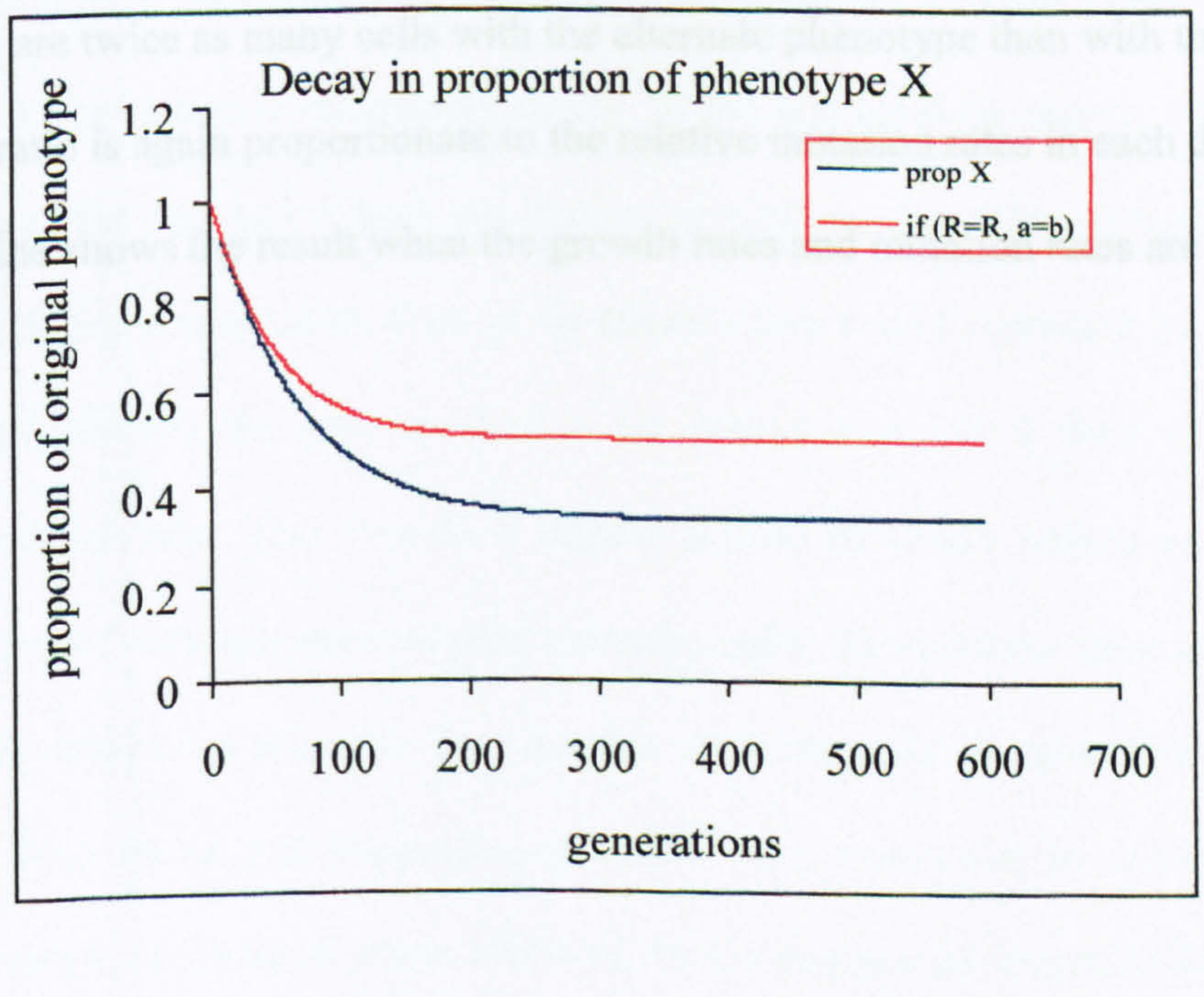


Figure 3.2. Simulations in which the effect of relative mutation rates on the final proportions of the phenotypes at equilibrium is shown. In each case a high mutation rate has been selected for illustrative purposes.

Figure 3.2a shows the result when the forward and back mutations are equal ($\alpha = ß = 0.01$). This results in an equilibrium population composition in which the proportions of the two phenotypes are equal. The ratio is proportionate to the relative mutation rates.

Figure 3.2b shows the result when the forward mutation is twice the back mutation rate ($\alpha = 0.01$ and $ß = 0.005$). This results in a final population composition (blue line) in which there are twice as many cells with the alternate phenotype than with the starting phenotype. The ratio is again proportionate to the relative mutation rates in each direction. A control red line shows the result when the growth rates and mutation rates are equal.

context of immune evasion because the host would be quickly exposed, and have opportunity to respond, to the full antigenic repertoire of the colonising population. Furthermore, genome analyses reveal that a single strain may have many phase variable characteristics as in *H. influenzae* (Hood *et al.*, 1996; Appendix 1), *H. pylori* (chapter 4), *N. meningitidis* (chapter 6), *T. pallidum* (chapter 5) and *C. jejuni* (Saunders & Jeffries – unpublished). If each of these were phase variable at a rate in the region of 10^{-2} then there would be insufficient stability for any clone to become adapted to a specific environmental niche without the accumulation of a large proportion of variant and potentially less fit phenotypes. Apart from the situation of colonisation, in which a small inoculum may benefit from rapid diversification in order to generate a clone with advantages in the initial stage of infection, these high rates would not be expected to provide a selective advantage for the population.

In contrast, figures 3.3.a, b and c show the effects of models when $\alpha = ß = 0.0025$, 0.0005 and 0.00005. This simulation demonstrates that in the absence of selection the composition of the population is remarkably stable. At the intermediate rate ($\alpha = ß = 0.0005$) more that 85% of the population have the starting phenotype after 600 generations. If one allows for a fairly rapid generation time of 30 minutes this would represent 12.5 days of growth and would involve the generation of a far greater population than could ever colonise an individual host. This reveals a central feature of phase variation. That although when compared with spontaneous point mutation rates the mutation rates are comparatively high, in the absence of selection, they are still associated with a reasonable degree of stability. In the time frame and bacterial population sizes that exist in colonisation and infection processes, the rates of phase variation, in the absence of fitness differences, do not lead to substantial changes in proportions of each phenotype in the bacterial population.

The situation is dramatically different when there are fitness differences between the phenotypes. Figures 3.4.a, b and c illustrate the influence of 1%, 10% and 50% reductions in the relative fitness of the starting phenotype relative to the alternate phenotype. A 1%

Figure 3.3a

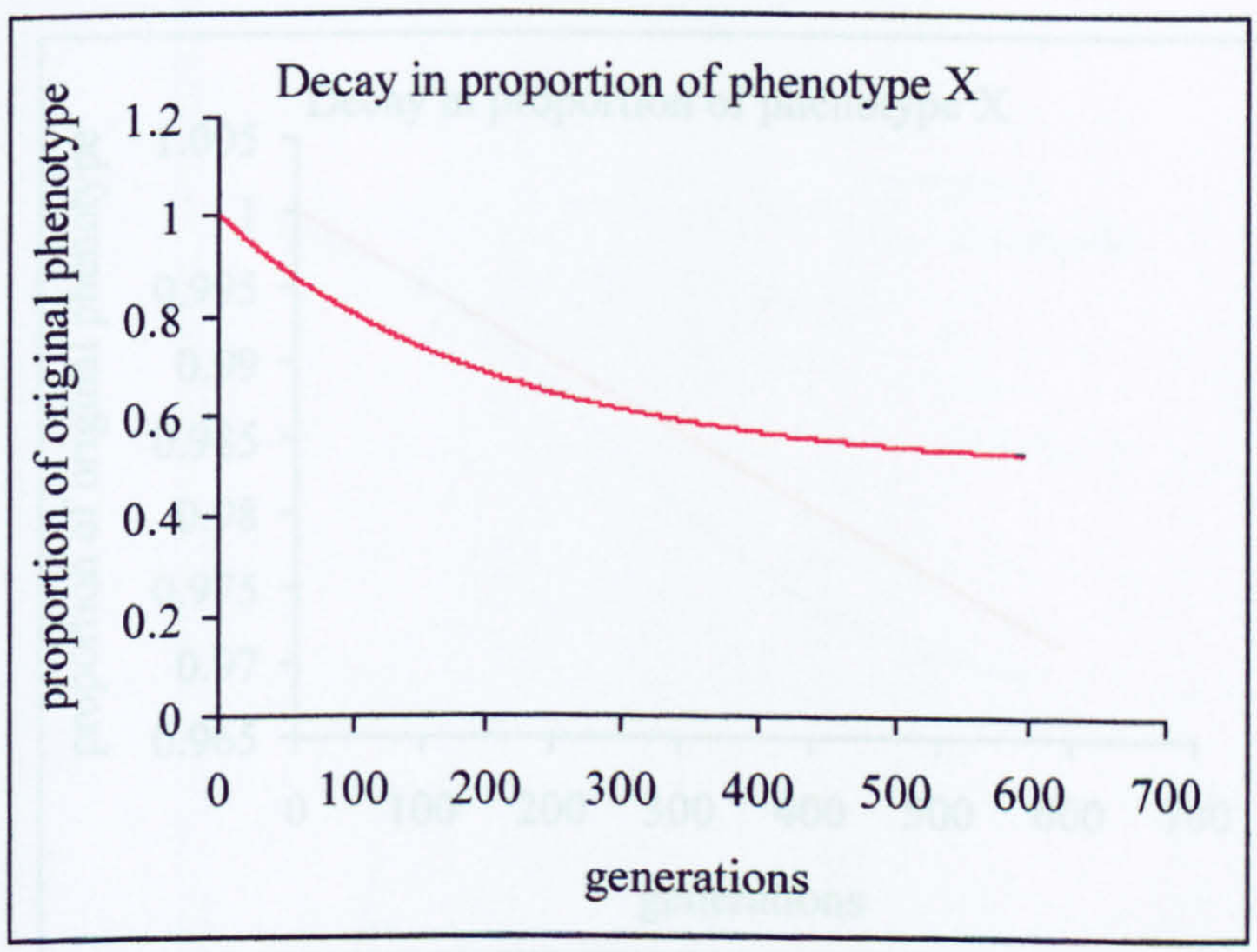


Figure 3.3b

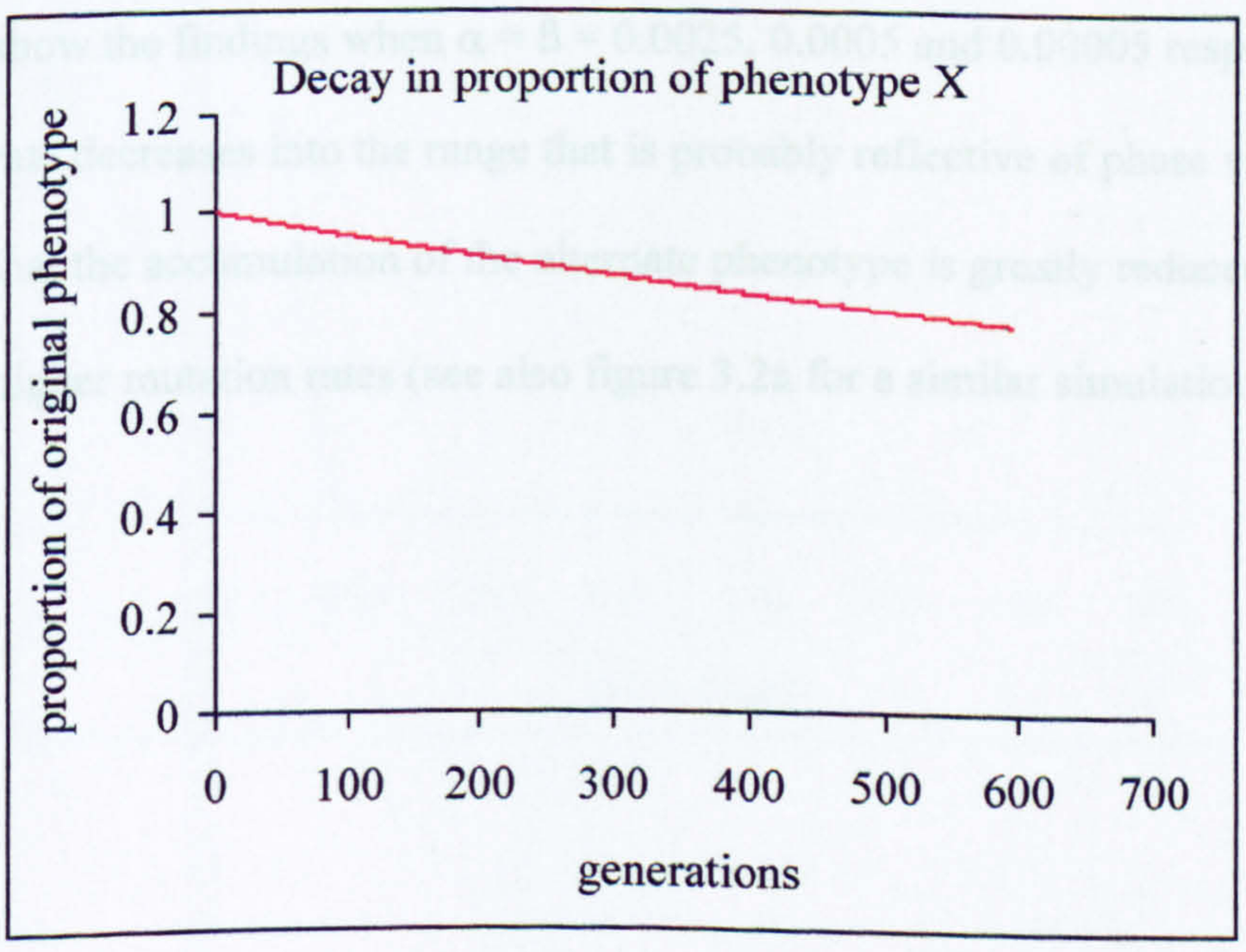


Figure 3.3.c

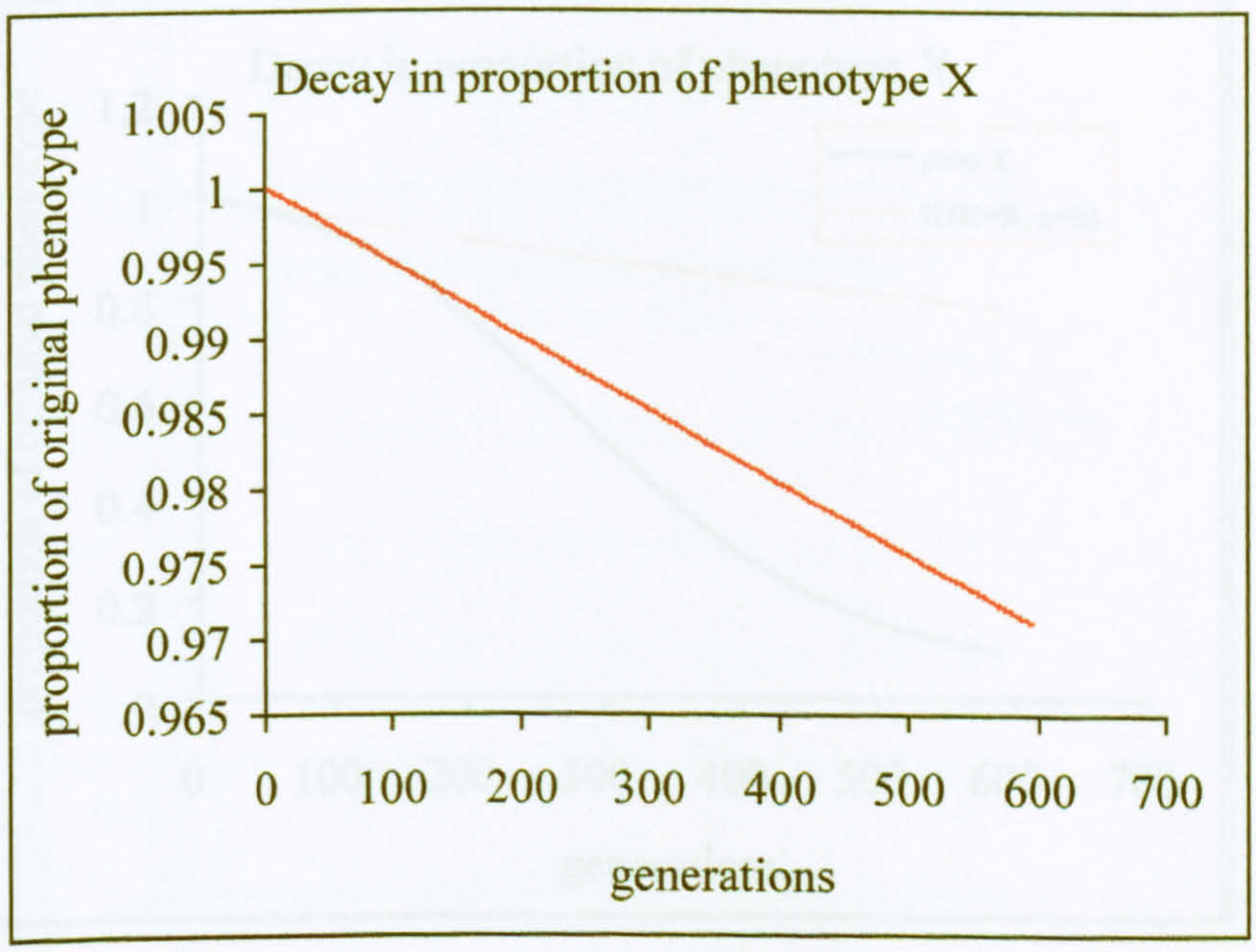


Figure 3.3. Simulations showing the effects of different mutation rates on population composition when the fitness of the alternate phenotypes are equal. Figures 3a, 3b and 3c show the findings when $\alpha = ß = 0.0025$, 0.0005 and 0.00005 respectively. As the mutation rate decreases into the range that is probably reflective of phase variation it can be seen that the accumulation of the alternate phenotype is greatly reduced when compared to higher mutation rates (see also figure 3.2a for a similar simulation when $\alpha = ß = 0.01$).

Figure 3.4.a

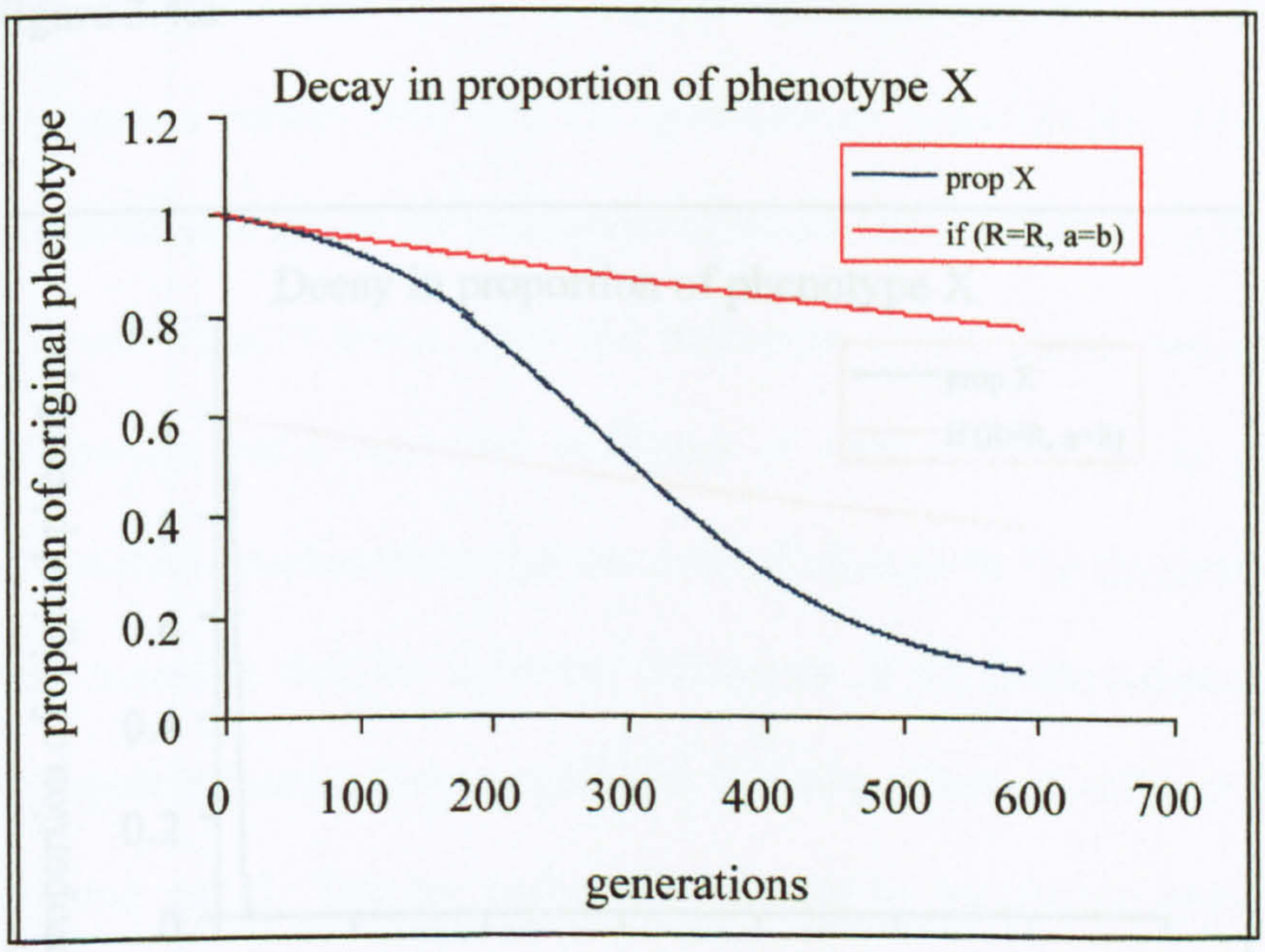


Figure 3.4.b

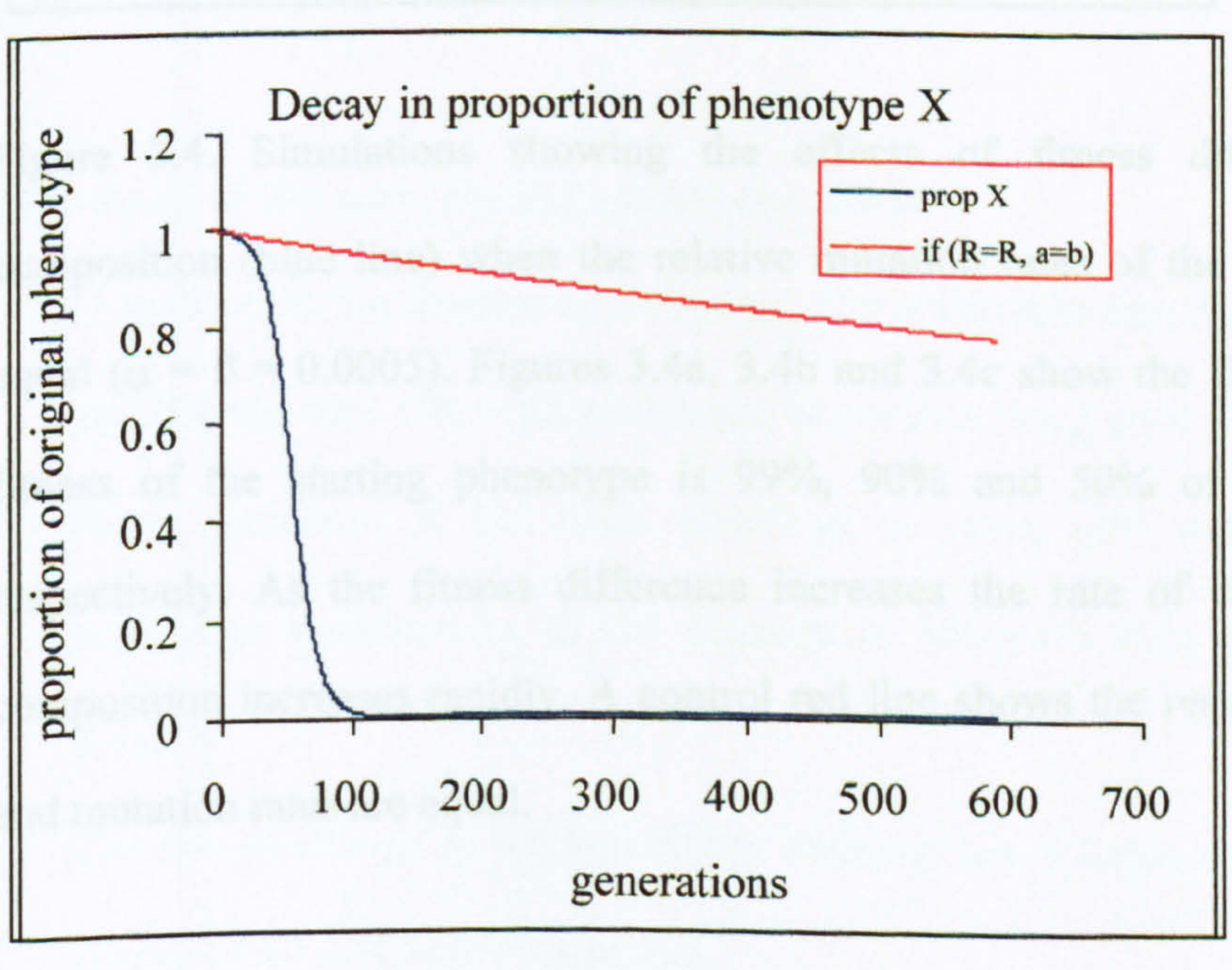


Figure 3.4.c

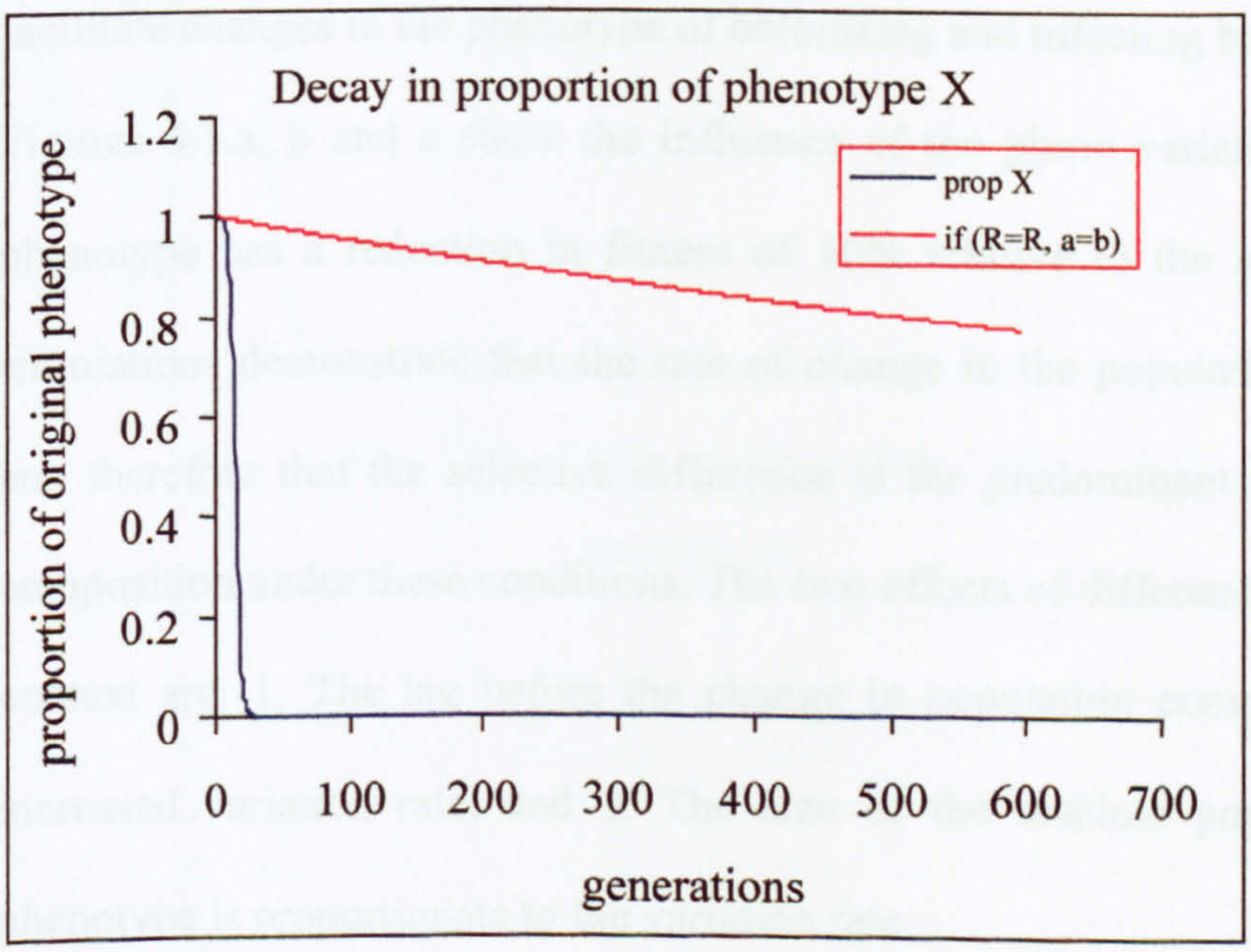


Figure 3.4. Simulations showing the effects of fitness differences on population composition (blue line) when the relative mutation rates of the alternate phenotypes are equal ($\alpha = ß = 0.0005$). Figures 3.4a, 3.4b and 3.4c show the findings when the relative fitness of the starting phenotype is 99%, 90% and 50% of the alternate phenotype respectively. As the fitness difference increases the rate of change of the population composition increases rapidly. A control red line shows the result when the growth rates and mutation rates are equal.

fitness difference results in a change in the predominant phenotype over 600 generations. Fitness differences of 10% and 50% (which may be modest in the presence of a specific immune response) result in almost complete replacement of the starting population phenotype within 100 and 20 generations respectively. This is sufficiently rapid to facilitate changes in the phenotype of colonising and infecting bacterial populations.

Figures 3.5.a, b and c show the influence of the phase variation rate when the starting phenotype has a reduction in fitness of 10% relative to the alternate phenotype. These simulations demonstrate that the rate of change in the population is similar in each case and therefore that the selective difference is the predominant determinant of population composition under these conditions. The two effects of differences in variation rates in this context are: 1. The lag before the change in population composition is decreased with increased variation rate, and 2. The size of the residual population with the starting phenotype is proportionate to the variation rate.

Summary of the main findings using the model:

1. Over time, in the absence of selection, phase variable populations will tend towards an equilibrium state that is proportionate to the switching rates for each phenotype.
2. Phase variation, at the rates observed *in vitro*, will result in stable phenotypic population composition in the absence of selection over biologically relevant time periods.
3. The rate of change in the population composition is largely determined by the relative fitness of the alternate phenotypes.
4. The time taken until changes in population phenotype composition is determined by the variation rate.
5. The proportion of the 'residual' phenotype increases with the phase variation rate.

Figure 3.5.a

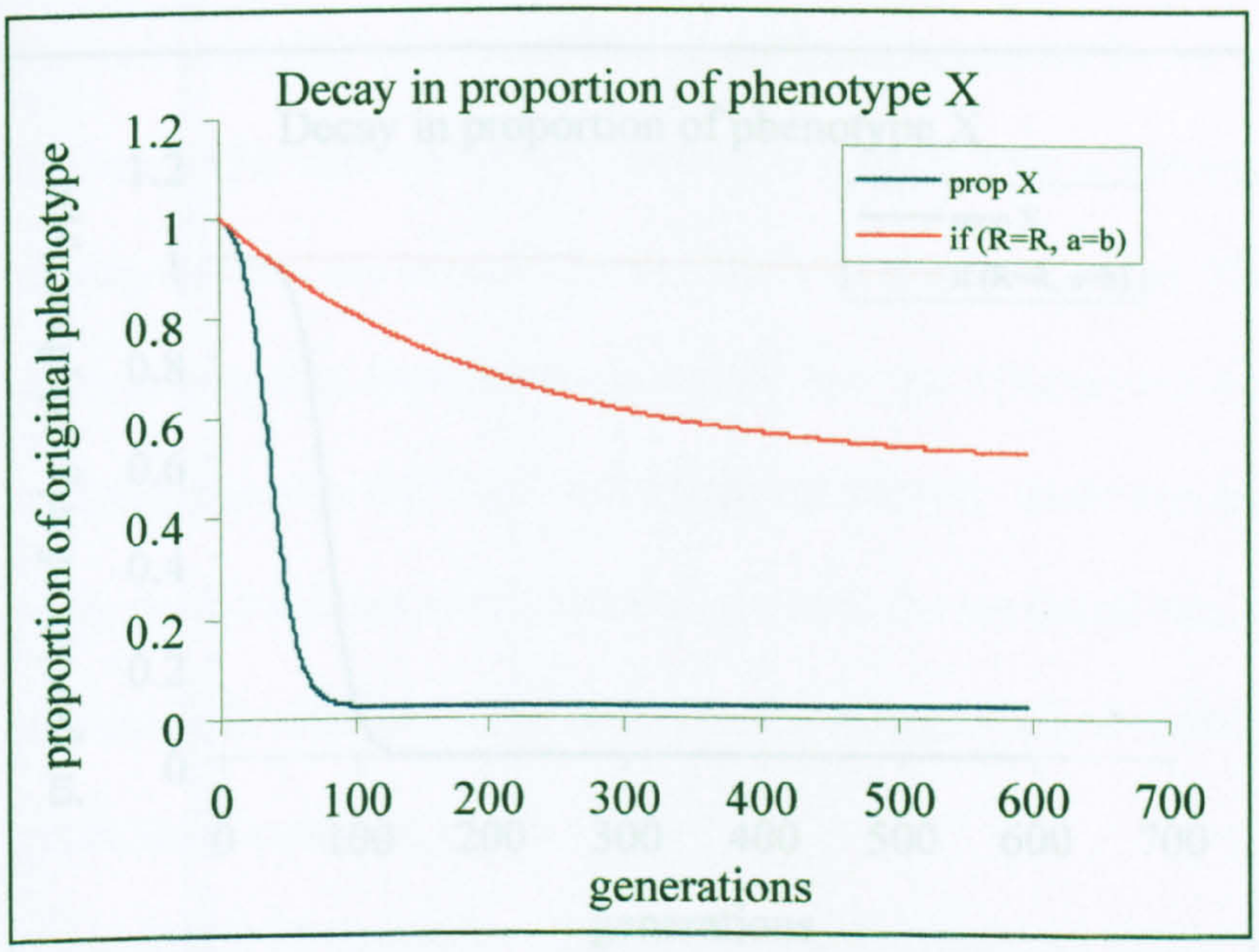


Figure 3.5.b

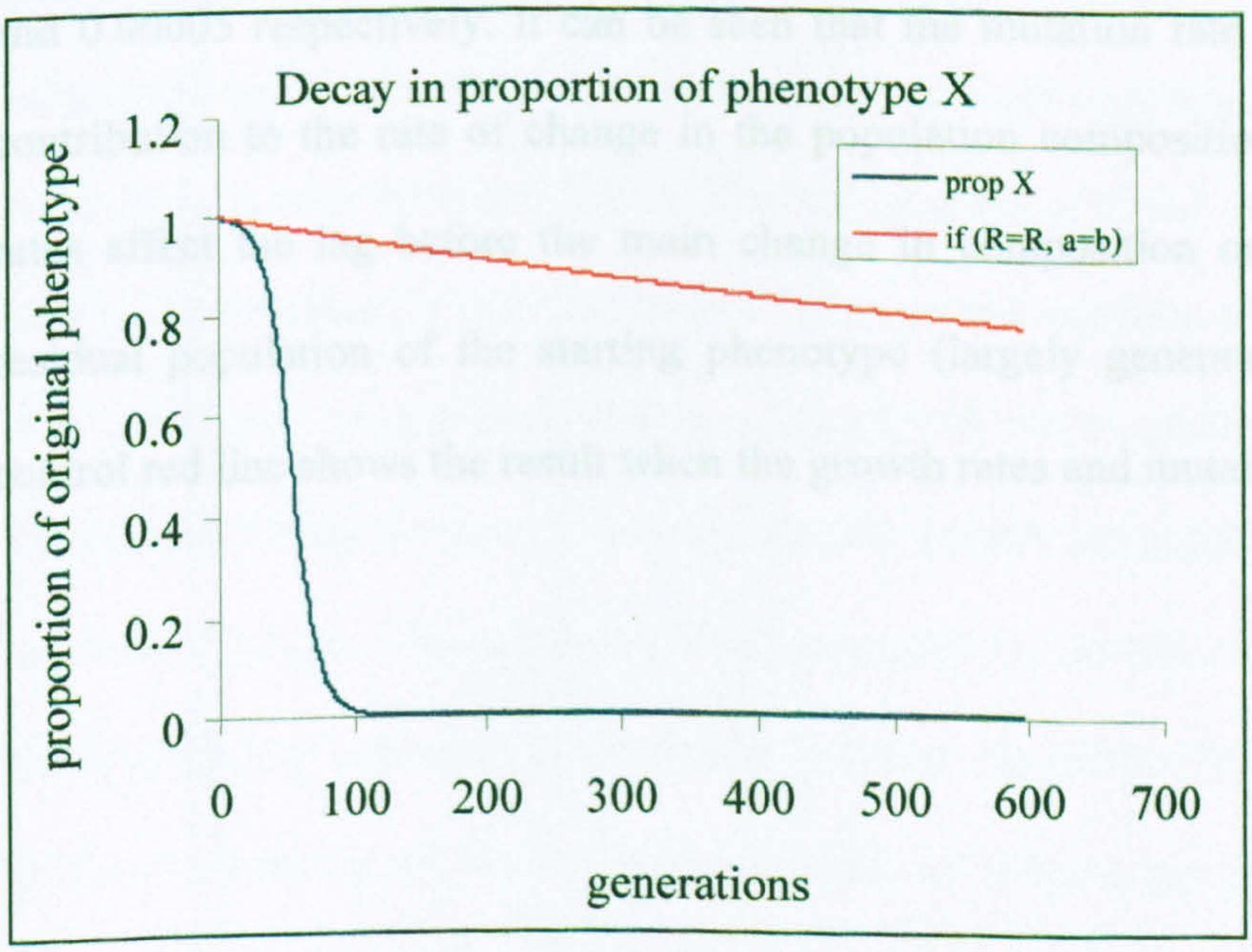


Figure 3.5.c

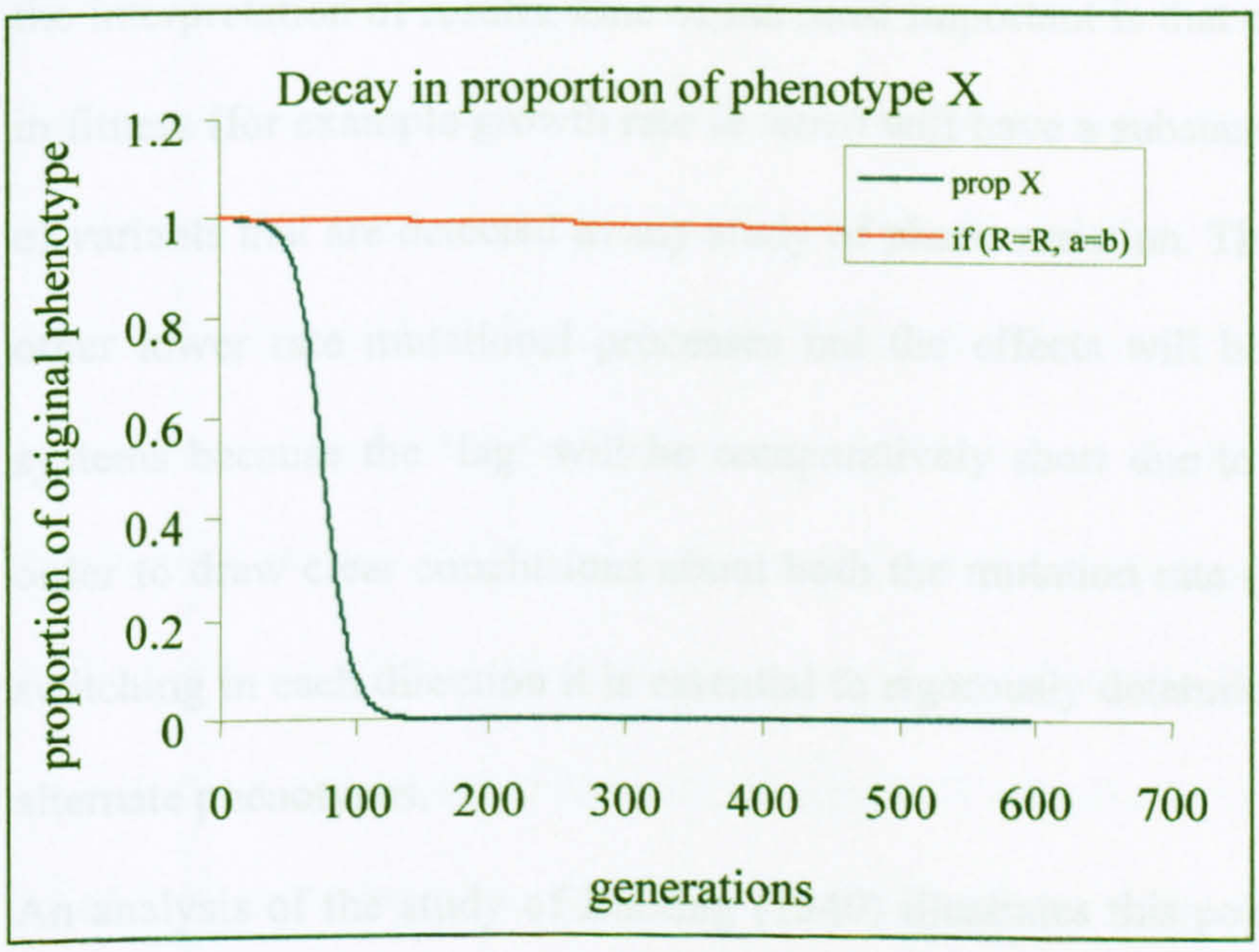


Figure 3.5. Simulations showing the effects of different mutation rates on population composition (blue line) when the fitness of the starting phenotype is 10% less than the alternate phenotype. Figures 3a, 3b and 3c show the findings when $\alpha = ß = 0.0025$, 0.0005 and 0.00005 respectively. It can be seen that the mutation rate makes a relatively minor contribution to the rate of change in the population composition. The different mutation rates affect the lag before the main change in composition occurs and the size of the residual population of the starting phenotype (largely generated by back mutation). A control red line shows the result when the growth rates and mutation rates are equal.

The findings of this study have several practical implications for experimental design and the interpretation of results. One of the most important is that even very minor differences in fitness (for example growth rate *in vitro*) will have a substantial effect on the proportion of variants that are detected in any study of phase variation. This is also true for studies of other lower rate mutational processes but the effects will be greater in phase variable systems because the 'lag' will be comparatively short due to the high mutation rate. In order to draw clear conclusions about both the mutation rate (μ) and the relative rates of switching in each direction it is essential to rigorously determine the relative fitness of the alternate phenotypes.

An analysis of the study of Bunting (1940) illustrates this point using an example which also demonstrates that the model allows interpretation of data in which the starting populations are mixed. In this study of colour variants in *Serratia marcescens* by Bunting (1940) it was recognised that phase varying populations approached an equilibrium state proportionate to the relative mutation rates. The populations were observed over a period of several days in serial subcultures such that a minor population was observed to increase from 22% to over 80% in 12 days. The author concluded that in their studied system there was an approximately 32:1 difference between the mutation rates in each direction. This was predicated on the basis that the phenotypes were equally fit in the *in vitro* culture conditions. Using the model with the starting population proportions and the approximate number of generations in Bunting's experiment then the fitness difference required to alter the population structure as described is only 1% (Rx = 0.99; Ry = 1). This fitness difference was not within the limits of detection of her experiment to show that the fitness of the two phenotypes were equal, so this experiment as it stands does not conclusively demonstrate the stated difference in the mutation rates. It is also noteworthy that the population changes are modelled equally well with rates of $\alpha = ß = 5 \times 10^{-4}$, 5×10^{-5} or 5×10^{-6}. This shows that under these circumstances, if the change in the population is affected by this level of selection, the rate of switching within this range is relatively unimportant.

Weiser and co-workers have progressed further than most other groups in the investigation of the effects of phase variation on fitness bacterial fitness in model systems and their findings can be considered in the light of the new model. The results of an experiment similar to the last simulation but in which the results were interpreted alternatively focussed upon the effect of phase variation of phosphorylcholine in *H. influenzae* in a mouse nasopharyngeal colonisation model (Weiser *et al.*, 1998a). If the data from this experiment are considered using the model (using: starting proportion X = 0.98 & Y = 0.02; $\alpha = \beta = 0.0005$; 1 generation per hour over 16 days = 384 generations) then what is presented as a dramatic difference in fitness is actually only approximately 1%. If the reversion rate is halved to represent the nature of the repeat mediated phase variation in this case then the fitness difference is still less than 1.25%. However, the growth rates of the two phenotypes were not compared to exclude a difference of this order of magnitude either on solid media or in the animal model, nor were the relative rates of mutation determined. The model demonstrates that in studies of this type it is essential to determine these parameters. Where this is not possible the assumptions must be clearly stated and the alternative interpretations of the data considered.

It also follows that if the relative mutation rates are known, or can be assumed to be equal, the relative fitness of the two phenotypes can be determined from the equilibrium population composition. This insight can be used to construct experiments to investigate the relative fitness advantages that different phenotypes confer on a population in infection models, under defined experimental conditions and even from samples obtained from natural infection. This means that rather than performing analyses on single samples from multiple individuals to investigate the fitness of a phenotypic combination (Weiser & Pan, 1998) one needs to collect multiple specimens from each individual patient or experimental model.

The predictions of population stability and compositional change based upon the model are consistent with observations from studies of phase variable systems. In *H. influenzae*

repeat loci associated with phase variable genes were found to be stable in prolonged daily subculture over a period of 16 days. The same loci were shown to display substantial heterogeneity from isolates obtained from two outbreaks of respiratory tract infection (van Belkum *et al.*, 1997a & b). There are 5 hexameric repeats present in *H. influenzae* strain Rd the stability of 3 of which have also been investigated. Of these three, 2 were found to be polymorphic whilst the third was apparently stable (van Belkum *et al.*, 1997a). If the location of these repeats is investigated it reveals that only the repeat that was not polymorphic was located in a sequence location where it would not affect the composition of the expressed protein. The 'stable' hexamer was the only one composed entirely of adenine and thymidines (TTAAAA), had the lowest melting temperature, and on this basis might have been expected to be the least stable of those studied. This reveals that repeat structures other than those associated with the on-off switching characteristic of phase variation can also be relatively hypermutable. It also supports the contention that it is the combined effects of variation and the generation of potentially selectable phenotypic differences that are required for changes in population composition to occur. *lic2* is associated with the production of a digalactoside lipopolysaccharide (LPS) structure associated with poor antibody mediated killing (Weiser & Pan, 1998) whilst *lic1* is associated with modification of LPS with phosphorylcholine which reduces CRP mediated killing (Weiser *et al.*, 1998a). Of the phase variable genes it is notable that the numbers of repeats associated with *lic2* and *lic1* are amongst the most variable in the epidemiologically related isolates. Thus genes which apparently adapt the bacterium for very different immunological pressures are amongst the most variable.

Phase variation has predominantly been investigated in human pathogens in which human studies are frequently not possible. The bovine pathogen *Haemophilus somnus* has phase variable LPS and causes respiratory tract and invasive disease in a similar fashion to its human specific counterpart. As seen with *H. influenzae* the LPS phenotype is stable over weeks of daily subculture. However, during infection in the natural host there are rapid

changes in the LPS of serial isolates. The appearance of variant phenotypes is associated with the generation of specific immune responses to the phenotype that is lost and the phenotypes occur sequentially as the animals are exposed and respond to each (Inzana *et al.*, 1992). This is the clearest demonstration of the fitness advantage to a colonising population of phase variation of LPS. It also illustrates apparent stability and then change in the absence and presence of specific selection pressures.

Despite the apparent stability of phase variable traits there are substantial theoretical difficulties in the investigation of the virulence of bacterial pathogens that possess multiple phase variable genes. At a rate of only $5x10^{-4}$ per generation per cell, under conditions in which there are no fitness differences between phenotypes, a population will accumulate approximately 1.3% variants for a single phenotype over the 25 to 30 generations during an overnight culture on solid medium. As described by the model this proportion will be much greater if the alternate phenotype is associated with fitness differences, and occasionally colonies will have greater variability due to the occurrence of mutation early in the generation of the colony (jackpot cultures). In an organism with 10 phase variable genes (there are potentially 65 in *N. meningitidis* (chapter 6)) this would result in the variation of one or more genes in 13% of the population. Furthermore the phenotypic composition of each colony will be different from any other. The greater the number of generations used to prepare inocula or occur during the subsequent experiment the greater the potential for diversification due to phase variation. The effect of very small fitness differences on population composition raises an additional potentially confounding factor. It cannot be assumed that the expression of one phase variable gene does not affect the increased or reduced fitness associated with the expression of others. For example, neisserial Opa proteins have been shown to interact with the variable portions of LPS, both of which are phase variable (Blake *et al.*, 1995). This is true of culture on artificial media as well as in models of infection. It is possible that variation in one gene will alter the fitness with respect to another and hence that the changes in one gene may be influenced

by the expression of others. In this way it is possible for independent stochastic switching processes to become effectively co-ordinated. Therefore, when observed, a change in a single phase variable gene cannot be assumed to be independent of other phenotypic changes.

Chapter 4

Putative phase variable genes associated with simple sequence repeats in the *Helicobacter pylori* genome.

4.1 Introduction

The publication of the complete, annotated genome sequence of a strain of *Helicobacter pylori* by the TIGR team (Tomb *et al.*, 1997) is a major contribution, facilitating research on this important pathogen. This chapter re-assesses and expands specifically upon their analysis and interpretation of simple nucleotide repeats and their potential importance to *H. pylori* biology. Previously, members of the Moxon group have demonstrated that multiple short tandem repeats can be used to identify a subset of genes (Hood *et al.*, 1996), designated contingency genes (Moxon *et al.*, 1994), that are hypermutable due to slippage within the DNA repeats. This slippage results in frequent shifting into and out of frame (relative to the translational start), leading to on - off switching of the associated gene products, and thereby phase or antigenic variation. Because of the types of gene in which these repeats are found, this potential for slippage should facilitate adaptation to changing host environments. These types of repetitive element have been used as markers with which to search genome libraries by Southern hybridisation in an attempt to identify novel genes involved in adaptability and host interaction (Peak *et al.*, 1996). The availability of whole genome sequence data allows this type of searching to be conducted '*in silico*'.

Tomb *et al.* identified seventeen putative phase-variable genes, based upon their finding of repeats in open reading frames of the *H. pylori* genome. A further 17 genes with upstream polyA or polyT tracts, in which gene expression might be regulated by variations in tract length, were also identified.

4.2 Methods and approach to whole genome analysis

This chapter describes the use of an integrated software system for prokaryotic genome analysis, which combines several strategies for identifying repetitive DNA. This has identified a number of additional putative phase-variable genes with internal repeats, their homologies, and their potential role in host:parasite interactions. It suggests reasons why it is unlikely that the previously described upstream repeats mediate phase variation.

The analysis system is based around ACEDB (Durbin & Mieg, 1991) an object-oriented database that has a graphical user interface (GUI) front end. ACEDB was originally developed to allow the visualisation of bibliographic, genetic and sequence data from the *C. elegans* genome sequence project, but was designed as a general database system which would allow easy and frequent extension and adaptation of the database schema, thus making it a very flexible genomic database. ACEDB facilitates the simultaneous graphical display of results from multiple analyses in the context of sequence location. Thus, independent analyses are seen in parallel (an example of this can be seen in figure 4) making interrelationships readily apparent. A second important aspect of this approach is the independent nature of the analyses. This means that the results of one dataset are not dependent upon the output of any other. This avoids potential problems related to the erroneous focusing on regions of interest, for example definition of predicted open reading frames.

A separate column is used to represent the output of the various analyses, within each column a box positioned relative to an index bar indicates the location of features. The vertical length indicates the span of the feature, while an indication of its significance is given by the width of the boxes. The width of each box is determined by a score, which is a function of some intrinsic property of the feature. This type of visualisation allows the interpretation of features such as the length and perfection of repeat elements. These characteristics can then be readily evaluated in the context of other features such as the

Figure 4

A sample screen from the analysis system showing a frame shift in the *H. pylori* FliP homologue. The title in the top of the frame identifies the contig being displayed. The black bar on the left hand side of the screen represents the whole of the contig and the green box upon it shows the proportion and location of the section being displayed in close-up. The adjacent green boxes display the results of the Array search, the largest of which represents the run of 9 Cs at the junction of the frame shift. The series of red bars displays the results of SIMPLEP. The long yellow bar and adjacent scale are for sequence location. The smaller yellow and the blue boxes show BLASTN and BLASTX homologies respectively. Each of these are to FliP genes from a variety of organisms. The pink boxes are ones that have been selected and the identity of the homology selected is displayed in the pale blue highlighted bar at the top of the screen. Only the 5' section was selected, the other two pink boxes display further 3' homology with the same gene. The very small yellow boxes and the horizontal black lines represent initiation and termination codons respectively, thus displaying open reading frames. The DNA sequence is displayed on the right hand side of the screen and the selected 5' portion of the FliP gene is highlighted. There are no boxes representing the homologies described by TIGR because this database was constructed prior to the availability of this annotation.

degree and span of identity that a neighbouring open reading frame has with other sequences. In addition, ACEDB includes several tools that facilitate the interpretation of features in the context of basic sequence characteristics such as identification of potential open reading frames and the location of initiation and termination codons, and each can be seen in a strand and frame specific context.

The complete *Helicobacter pylori* genome sequence in Fasta/Pearson format was downloaded by ftp from the TIGR ftp site (URL:ftp//ftp.tigr.org/pub/data/h_pylori/GHP.1con.Z). This unannotated sequence data was partitioned using a PERL script into shorter 101kB contigs, with each contig overlapping the previous contig by 1kB. These contigs were in turn processed by a series of nested scripts, that were designed to automate a large proportion of the computational analyses.

One of the strengths of our system is that while it can be used in conjunction with the published sequence annotation it is essentially independent of this annotation. This allows us to analyse preliminary unannotated sequence data in much the same way as fully annotated genomes, making the analysis independent of any mistakes or assumptions in the published annotation. However, this independence comes at the price of carrying out a separate analysis of open reading frames and searches for sequence similarity.

Sequence similarity was identified using BLASTN (searches a DNA database with a DNA query sequence) and BLASTX (searches a protein database with a DNA query sequence conceptually translated into all 6 reading frames), both programs are from the BLAST program suite (Altschul *et al.*, 1990). The databases searched were specifically constructed using the sequence retrieval system (SRS) (Etzold and Argos, 1993), such that all metazoan and higher sequences were excluded (to reduce processing time), and to separate sequences from Helicobacter species; this overcomes the problem of matches against "self" overwhelming other significant but weaker matches. The DNA databases were constructed

from a non-redundant set of sequences from the EMBL and GenBank databases (Emmert *et al.*, 1994), while the protein databases were constructed from a non-redundant set of sequences from the SwissProt (Bairoch & Apweller, 1997), PIR (George *et al.*, 1994) and TREMBL (Bairoch & Apweller, 1997) protein databases. Each of the contigs was searched against a "self" and "non-self" protein database. The resulting output files from these searches were further processed to remove weak matches and to convert the output files into the format required by ACEDB.

In this study, five separate programs: array, tandem, perfect tandem, SIMPLEX and SIMPLEP, were used to identify and locate repeat sequences. Array (provided by J. Hancock - MRC Clinical Sciences Centre) was used to identify perfect repeats with component motifs of between 1 and 10 bases, which were repeated at least five times. Tandem and quicktandem, both part of EGCG (URLhttp://www.sanger.ac.uk/Software/EGCG) were used to identify all tandem repeats, tandem was used to rigorously search for all repeats up to 150 bp in length, that had at least 80% sequence similarity. The arbitrary 150 bp limit was chosen due to constraints on processing time, but quicktandem, a less rigorous but faster program was also used to search for other tandem repeats between 150 bp and 1kb in length.

Local repetitivity was analysed using SIMPLEX and SIMPLEP (Saunders, Peden & Moxon - in preparation) which are both based upon SIMPLE34 (Hancock & Armstrong, 1994). Both programs identify repetitivity by estimating if a word occurs significantly more often within a local window around the word than with the frequency with which it occurred within a window of the same size in a random sequence. The random sequences were constructed by shuffling the dinucleotide pairs of the original sequence. SIMPLEX was a basic modification of SIMPLE34 such that it could analyse much larger sequences (including whole genomes). SIMPLEP was however a complete rewrite of SIMPLE34 from FORTRAN into PERL. SIMPLEP increased the maximum length of the word used to

identify repetitivity; it also doubled the size of the window to 128 bp. Most importantly it altered the method for identifying repetitivity from the purely word based method of SIMPLE34 and SIMPLEX to one that identified a repetitive local region. This was achieved by not estimating the significance of each word in isolation, but by considering the frequencies of neighbouring words; in this analysis we considered neighbouring words to be the 10 nearest words. This change increased the sensitivity of the search such that the limit for significance had to be increased from 3 to 6 standard deviations.

The graphical display was used to evaluate the context of all homopolymeric tracts and dinucleotide repeats of 10 bases or longer. In addition, all repeats located at the junctions of two open reading frames were examined. In order to ensure that no repeats were omitted from the analysis using the graphical interface, the results of the array and tandem searches were exported in a tabular format and this was compared with the repeat elements identified using the graphical interface. When homologies were not identified by the system in regions where putative open reading frames were potentially associated with functional repeats the sequence was extracted and separate BLASTN and BLASTX searches were performed (Altschul *et al.*, 1990).

As described above, repeats were displayed graphically in the context of putative homologues identified by DNA and amino acid sequence identity, and predicted open reading frames. We have used this system to analyse the genome of *H. pylori* and to compare it with other completed, publicly available bacterial genomes for the types of repeat that were suggested to have function by Tomb *et al.* This integrated approach, as described above, has several advantages over other search methods:

1. Programs that identify all repetitive DNA sequences, when used in isolation, create a practical problem of how to select those repeats likely to have biological function. To constrain the number of 'hits' whilst maximising potentially useful 'finds', Tomb *et al.* used a search program with an arbitrary cut off (e.g. 10 nucleotides in a homopolymeric

tract). Our system allows greater sensitivity because the identified motifs are seen in context and large numbers of them can be evaluated at one time. For example, the homologue of the gene for a flagellar protein (*fliP*) has a repeat of 9Cs associated with a frame shift so that its gene product would not be expressed. This gene is likely to be involved in flagellar assembly so that this would disrupt flagellar synthesis and motility completely. Using a cut-off of 10 nucleotides, Tomb *et al.* excluded this gene and its repeats from their search.

2. The analysis and annotation of Tomb *et al.* was based upon identification of putative open reading frames. In contrast, our system analyses homologies, repeats and open reading frames independently using the genome sequence as a whole. This identifies and allows for frame shifts due to the length of the repeat elements as well as any (rare) sequencing errors. As an example, we identified a homologue of an adenine specific methyltransferase which had been divided into three segments by two homopolymeric tracts. The first two segments had been identified as separate open reading frames (HP1353 and HP1354), the third was not attributed. To date we are not aware of any gene that is phase variable due to the presence of two independently acting homopolymeric tracts.

3. Our system also provides information to help interpret the significance of intergenic repeats located between genes. For example, a repeat of 8 Cs occurring 7 bases before a start codon, as in the case of the *secD* (HP1550) homologue, would be less likely to interfere with promoter activity than a run of 12 Cs located 75 bases upstream of a start site, as is the case for the homologue of a methyl accepting chemotaxis protein (HP0103).

4. Our approach facilitates identification of related genes in different organisms associated with particular motifs. For example, the strong homology between methylases associated with a homopolymeric repeat in *H. pylori* and tetrameric repeats in *H. influenzae*, respectively.

5. The data files used to generate the graphical output are retained and can be used to analyse the repeats present in the genome as a whole.

4.3 Results and discussion

4.3.1 Comparative analysis of repeats present in *H. pylori* with those in other species

Comparison with other genomes (*Escherichia coli*, *Haemophilus influenzae*, *Methanococcus janaschii*, *Borrelia burgdorferi*, *Mycoplasma genitalium* and *Mycoplasma pneumoniae*) revealed that the prevalence of homopolymeric and dinucleotide repeats in *H. pylori* is strikingly different from those of these other published genomes. In *H. pylori*, homopolymeric and dinucleotide homopurine:homopyrimidine tracts are comparatively frequent. If the sequences of the above listed complete genome sequences are combined then *H. pylori* has 18 of 23 (78%, $p<0.0001$) poly(A/T) runs of >11 bases, 23 of 28 (82%, $p<0.0001$) poly(C/G) runs of >8 bases, and 9 of 10 (90%, $p<0.0001$) poly(GA/TC) runs of >5 repeats when compared with all of these available sequences. The *H. pylori* genome represents 14% of the total nucleotides and does not lie at either extreme of G:C content (39%).

4.3.2 Identification of putative phase variable genes

This analysis identified 27 candidate genes that are likely to be phase variable owing to the presence of simple sequence repeats (Table 4.1), of which 26 have intragenic repeats.

Table 4.1 Describing the candidate phase variable genes identified in the *H. pylori* genome sequence.

Repeat	F/S	Homologies	TIGR HP number
LPS biosynthesis			
(C)13	-	Alpha (1-3) fucosyltransferase	0651
(C)13	-	Alpha (1-3) fucosyltransferase	0379
(G)14	+	Alpha (1-2) fucosyltransferase	0093 & 0094
(C)13	+	Lex2B	0619
(AG)11	+	RfaJ / glycosyltransferase	0208
Cell surface associated proteins			
(C)9	+	FliP	0684 & 0685
(GA)9	+	OM adherence protein associated protein	1417
(C)15	+	Streptococcal M protein	0058
(CT)6	-	outer membrane protein	0638
(GA)5	-	alginate O-actetylation protein	0855
(T)9	-	heme binding lipoprotein / transport protein	0298
(C)12	PRO	methyl accepting chemotaxis protein	0103
(CT)11	-	outer membrane protein / adhesin	0896
(CT)8	+	outer membrane protein / adhesin	0722
(CT)6	+	outer membrane protein / adhesin	0725
(A)10	-	ABC transporter / secretion protein	1206
DNA restriction / modification system			
(C)12 & (C)15	+/+	adenine specific methyltransferase	1353 & 1354
(C)10	+	R/M methyltransferase	1369 & 1370
(C)14	-	restriction enzyme beta subunit	1471
(G)12	+	adenine specific methyltransferase	1522
(G)15	-	type I restriction enzyme R protein	0464
(CT)5	-	cytosine specific methyltransferase	0051
Hypothetical open reading frames without identified homologies			
(T)8	+		0586 & 0585
(G)12	-		0217
(AG)9	+		0744
(AT)5	-		0211
(G)9	+		0335

F/S + = presence of frame shift in published sequence; PRO indicates repeat located within promoter region

The 26 candidates include all 17 genes previously identified by the presence of intragenic repeats by Tomb *et al.* In 5 instances it was possible to merge what were previously considered to be distinct adjacent reading frames. Inclusion of the intervening sequence has allowed new homologies to be identified. One particularly instructive example is what we

have now identified as an alpha(1-2)fucosyltransferase homologue. Identification of this putative gene required integration of what were previously 2 distinct putative genes and the previously unassigned intervening sequence before this homology could be determined.

When compared with the other available genomes, the greatest difference in the frequency of homopurine:homopyrimidine tract lengths in *H. pylori* occurs at >8 bases. This is the maximum length that is corrected by DNA polymerase proof reading in yeast (Tran *et al.*, 1997) and thus mismatch repair becomes of greater importance for the stability of longer tracts. We assume that bacterial DNA polymerase has similar length dependence. Perhaps the greater prevalence of longer homopolymeric tracts in *H. pylori* reflects differences in the efficiency of mismatch repair, an inference consistent with the reported absence from *H. pylori* of MutH, MutL and components of the SOS system (Tomb *et al.*, 1997). The prevalence of these repeats, especially polyA or polyT, and the proportion of them between putative genes, suggests that considerable caution must be used in ascribing phase variable functions based on tract length alone. To the contrary, we suggest that the presence of the intergenic polyA or polyT tracts is unlikely to be associated with phase variation for the 17 genes identified by these markers (Tomb *et al.*, 1997). If they have any function at all when located intergenically, this function may be related to transcription termination or promoter enhancement (Ellinger *et al.*, 1994).

The potentially phase variable genes identified in *H. pylori* can be divided into three groups: LPS biosynthesis, cell surface associated proteins, and DNA restriction / modification systems. Previously identified phase variable genes associated with repeats in *H. influenzae* (Hood *et al.*, 1996) fall into the same groups, as do those of other bacterial pathogens. The published genes with which the repeat associated genes share sequence identity give an indication of the functions from which Helicobacter derives a selective advantage through varying its phenotype during host:parasite interactions.

LPS is a polysaccharide cell surface component that is intimately associated with the host:parasite interface and has the potential to influence this interaction and also presents epitopes for immune responses. Phase variation of LPS structures has been described in other bacterial species. In *H. influenzae*, 5 distinct biosynthetic loci are thought to be associated with high frequency phase variation of surface exposed epitopes (Weiser *et al.*, 1989b; Hood *et al.*, 1996; Jarosik & Hansen 1994), each of which is associated with intragenic tetrameric repeats. In *Neisseria spp.* genes involved in the synthesis of superficial LPS structures which can act as substrates for further substitutions such as sialylation are phase variable and are associated with intragenic homopolymeric tracts (Gotschlich 1994; Jennings *et al.*, 1995). The variable loci in combination provide the cell with the ability to generate several LPS phenotypes providing opportunity for immune evasion and altered adhesion characteristics, whilst substitutions can further alter immunological properties and resistance to opsonophagocytosis (Kim *et al.*, 1992; van Putten, 1993). The presence of a similar number of repeat-associated loci in *H. pylori* compared with *H. influenzae* suggests a similar potential for flexibility of LPS structure. This includes a homologue of related function to a gene in *H. influenzae*, *lex2*, an LPS biosynthetic gene in *H. influenzae*, known to be phase variable owing to multiple repeats of the tetramer GCAA (Jarosik & Hansen 1994; Hood *et al.*, 1996). Furthermore, the identification of a repeat associated with alpha(1-2)fucosyltransferase, in addition to the two alpha(1-3)fucosyltransferases, helps to complete our knowledge of a recognised virulence determinant of *H. pylori*. *H. pylori* is known to phase variably express human Lewis X and Y epitopes on its LPS (Appelmelk *et al.*, 1999). This new homologue would be required for Lewis Y synthesis, and the presence of repeats in this gene, as well as the two alpha(1-3)fucosyltransferases, provides a possible mechanism whereby expression of this repertoire of surface structures is varied. This permits the organism to evade immune

responses by host mimicry (Appelmelk *et al.*, 1997) and has been reported to influence adhesion to the gastric mucosa (Boren *et al.*, 1993).

The homologue of HP 0855 is a protein required for acetylation of surface alginate, an exopolysaccharide with similarities to bacterial capsule. Overproduction of alginate is a putative virulence determinant of *Ps. aeruginosa* where it is characteristic of strains causing chronic lung infections in cystic fibrosis (Deretic *et al.*, 1995). Acetylation affects the properties of alginate including viscosity and binding of calcium ions (Skjak-Braek *et al.*, 1989) and may be important in the virulence associated functions of the exopolymer. Variation in acetylation of other polysaccharide structures has been demonstrated and has been linked with resistance of the polysaccharide to enzymatic degradation and antigenic variation. An example of this is the phase variation of O-acetylation of the polysialic acid K1 capsule of *Esch. coli* (Orskov *et al.*, 1979) which is mediated by variable expression of polysialic acid O-acetyltransferase (Higa *et al.*, 1988). Although *H. pylori* lacks a capsule, its LPS is known to be a complex, variable structure. These findings suggest that fine structural analysis of the structures should be done to determine whether O-acetylation occurs.

Variation of expression of extended surface structures important in motility and adherence such as pili, fimbriae, and flagella is a common phenomenon in bacterial pathogens. This variability is achieved by a number of diverse mechanisms. In *Esch. coli* fimbrial variation is achieved by a site-specific recombination event, the rate and direction of which is influenced by binding of IHF, that alters promoter direction and controls the expression of *fimA* (Abraham *et al.*, 1985; Higgins *et al.*, 1988). In *Neisseria spp.* altered pilus expression is achieved by recombination between expressed and silent pilus genes (Haas & Meyer, 1986). In *H. influenzae* the expression of fimbriae is controlled by alterations in the length of a dinucleotide repeat located within the divergent promoters of *hifA* and *hifB* which encode the structural sub-unit and export proteins respectively (van Hamm *et al.*,

1993). We are not aware of reports of variable motility in *H. pylori* but altered expression of FliP, a component of the basal structure of flagella involved in protein export (MacNab, 1996) within which the repeat is located, would be expected to result in a clean on-off switch of this surface structure.

Phase variable surface proteins that are not part of extended structures are also recognised as exemplified by the neisserial Opa and Opc proteins. Adaptive variation of Opa proteins is observed during infection with *N. gonorrhoeae* (Jerse *et al.*, 1994) and alterations in expression have been shown to alter tropisms for and interactions with epithelial, endothelial and monocytic cells (Kupsch *et al.*, 1993). The presence of putative phase variable surface proteins with homology to adhesins, flagellar components, surface substitution mechanisms and LPS suggest that *H. pylori* has a large and flexible repertoire of surface structures with which it adapts to its niches within the host and evades immune defences that are probably crucial to its characteristic persistence.

The repeat associated methylase in *H. influenzae* strain Rd (Hood *et al.*, 1996) was the closest homologue to one of those identified in *H. pylori*, and a repeat associated modification enzyme has also been described in *N. gonorrhoeae* (Belland *et al.*, 1996). Discerning a selective advantage resulting from phase variation of this type of gene is problematic. Inactivation of a methylase involved in restriction modification might be expected to lead to the death of the unmethylated progeny cells within which the restriction enzyme persists. A phase variable restriction system component could therefore be a remarkably 'unselfish gene'. It is possible that a population of transformable bacteria might derive a benefit from a continuous process of 'bacterial suicide' by a proportion of that population. However, other possible roles for these putative phase variable genes in this group of organisms, perhaps related to DNA metabolism, that are both naturally transformable and use slippage-like processes in phase variation should be considered. In

either context, the significance of the greater number of these mechanisms observed in *H. pylori* than have been observed in other bacterial species needs to be considered.

Whilst the repeat associated genes identified in this analysis of *H. pylori* have functional relationships to those seen in other organisms, the constituent nucleotides and the location of the repeats is different, indicating evolutionary convergence of similar mechanisms for phase variation. *H. pylori* provides a further indication of a key role for contingency genes in the host:parasite interaction and this illustrates the utility of this type of *in silico* analysis of genome sequence data in the search for genes and corresponding structures that are important in interactions of the microbe with the host.

4.4 Subsequent work supporting the results of this analysis

Following this study there are two groups whose work has confirmed predictions from this paper. In two papers Ben Appelmelk's group have demonstrated phase variability mediated by variation in the homopolymeric tracts located within the LPS genes (Appelmelk *et al.*, 1998 & 1999). Secondly, the sequencing of a second strain of *H. pylori* has revealed polymorphisms within several of the repeats highlighted in this analysis that would be associated with altered expression of the associated genes (Alm *et al.*, 1999).

Chapter 5

An investigation of simple sequence repeats in *Treponema pallidum* and the identification of repeat associated potentially phase-variable genes

5.1 Introduction

T. pallidum is an obligate human parasite that cannot be continuously cultured *in vitro* (Norris, 1982). The lack of a culture method has impeded the study of this organism and has resulted in *T. pallidum* being one of the most poorly understood species associated with human infection (Norris *et al.*, 1993). No known virulence factors have been identified and the organism is thought to have very few surface proteins (Radolf *et al.*, 1989; Walker *et al.*, 1991 ; Cox *et al.*, 1992). The natural history of primary, secondary and tertiary syphilis (Tramont 1995; Singh & Romanowski, 1999) indicate critical host interactions, for example with cardiovascular and neurological tissues. The primary clinical infection is localised to the site of inoculation but the organism is disseminated throughout the body at an early stage. In addition to the local features of infection, which involve inducing the formation of a papule and subsequently an ulcer, the invasive disease is characterised by interactions with lymphoid tissue and tropisms for vascular and neural tissues and a variety of characteristic lesions. Interactions with a variety of cell surfaces at deep sites in disseminated disease, the chronic nature of the infection and the ability to evade the immune response, are all consistent with the capacity to express phenotypes which fit the organism to the various conditions that it encounters.

Analysis of the recently completed genome sequence of *Treponema pallidum* (Fraser *et al.*, 1998) has presented an opportunity to gain insights into the biology of this important human pathogen. The genome sequence was interrogated for repeat associated ORFs as a hypothesis generating exercise to identify potentially host interactive phase-variable genes. This inherently speculative approach has highlighted systems involved with motility,

latency and the variation of a family of surface proteins, as well as identifying a number of genes with no known function as of potential importance in virulence.

5.2 Methods

The *T. pallidum* complete genome sequence was processed and interrogated using the previously described analysis software (Sections 5 and 6; Saunders *et al.*, 1998). The repeats present within the genome were analysed using: **ARRAY** (a program that searches for all perfect repeats with 5 or more copies of the component motif, in which motifs are less than 10 bases in length), **SIMPLEP** (an in-house program that uses a probabilistic method to identify sections of genome that are more repetitive than would be expected by chance), **TANDEM** (10 to 250 base imperfect direct repeats) and **QUICKTANDEM** (250 to 1000 base direct repeats). The whole genome was searched for sequence identity using **BLASTN** and **BLASTX** against archaeal, bacterial and lower eukaryotic sequences (i.e. excluding metazoa), after the databases had been divided into 'self' (*Treponema*) and 'non-self' (non-*Treponema*) sections. The complete genome and coding regions, as described in the TIGR annotation, were downloaded from The Institute for Genomic Research (TIGR) ftp site (ftp://ftp.tigr.org/pub/data/t_pallidum/) and were used in the database for the 'self' BLASTX analysis. For reasons of practicality the complete genome sequence was split into 101kB sections (with 1kB overlaps) and the results of these analyses were loaded into ACEDB.

Initially the data was examined with the aid of the ACEDB graphical interface. Repeats of 10 or more bases in length, and those associated with frame shift changes in ORFs, were initially considered to be potentially functionally unstable. Several repeats associated with frame-shifts in some potentially phase variable genes were homopolymeric tracts that were shorter than 10 bases in length. Based upon this observation, all C or G repeats of greater than 7 bases and A or T repeats of greater than 8 bases were also examined. The results of

the **ARRAY** and **TANDEM** searches were exported and manually cross-checked with those found using the graphical interface.

The frequency of each homopolymeric tract of A or T and C or G of 2 to 12 bases in length was used to determine the expected numbers of each homopolymeric tract length using high order Markov chains (Cox & Miller, 1965). For comparative purposes, the frequency of each homopolymeric tract of length L was predicted from the frequency of tract length L-1 with the hypothesis that the Lth base was determined by random distribution. That is, if the frequency of TTTT is 9521 copies in the genome, then the "expected frequency from previous" frequency for TTTTT would be 9521 * 0.237 = 2256, where 0.237 is probability of a thymidine.

5.3 Results and Discussion

Those repeats that may mediate phase-variation and the open reading frames with which they might be functionally associated are listed in the Table 5.1.

Table 5.1: Putative repeat associated phase-variable genes in the *T. pallidum* genome sequence.

Repeat	Frame shift	Gene similarities	Rating	TIGR number
Surface structures				
(G)9	pro	Flagellar motor switch protein 1 (*fliG-1*) and a putative hemolysin[1]	S	TP0026 & TP0027
(A)9	(+)	Flagellar filament cap protein (*fliD*)[2]	S	TP0872
(A)9	-	Flagellar-associated GTP-binding protein (*flhF*)	M	TP0713
(A)9	-	Flagellar basal-body rod protein (*flgG-2*)	M	TP0961
(G)9	-	Glycerophosphoryldiester phosphodiesterase (*glpQ*)	S	TP0257
(G)8	-	Rod shape determining protein (*mreB*)[3]	M	TP0497
(T)10	(+)	Rod shape determining protein (*mreC*)[3]	S	TP0498
(A)9	-	Penicillin binding protein (*pbp-1*)[3]	M	TP0500
(A)10	-	Oligopeptide ABC transporter, periplasmic binding protein (*oppA*)[4]	S	TP0585
(T)9	-	Thiamine ABC transporter, permease protein[5]	M	TP0143
DNA metabolism				
(A)10 &	-	Exonuclease (*sbcC*)	S	TP0627

(A)9				
(A)10	-	Excinuclease ABC, subunit A (*uvrA*)	S	TP0514
(A)9 & (A)9	-	Chromosomal replication initiator protein (dnaA)	M	TP0001
(A)9	-	DNA ligase (*lig*)	M	TP0634
General metabolism				
(TG)5.A .(G)10	pro	Preprotein translocase subunit (*secA*)	S	TP0379
(G)8	-	Long-chain-fatty-acid CoA ligase	M	TP0145
2 x (G)8	-	Heat-shock protein 70 (*dnaK*)	M	TP0216
(G)8	-	Primosomal protein N (*priA*)	M	TP0230
(TG)6	-	Pheromone shutdown protein (*traB*)[6]	M	TP0953
(G)6	+	Pyrophosphate-fructose 6-phosphate 1-phosphotransferase (*pfk*)[7]	M	TP0108[8]
ORFs of unknown function				
(G)11	-	Hypothetical	S	TP0127
(G10)	+	Hypothetical	S	TP0618 & TP0617
(C)13	-	Hypothetical[9]	S	TP0347
(G)9	-	Hypothetical	S	TP0479
(G)9	-	Hypothetical	S	TP0697
(G)9	+	Hypothetical[10]	S	(TP0135)
(G)10	-	Hypothetical	S	TP0126
(G)9	+	Hypothetical	S	TP0066 & TP0067
(G)10	+	Hypothetical	S	Not annotated[11]
(G)13	+	Hypothetical[12]	S	TP0382
(G)11	+	Hypothetical	S	TP0859 & TP0860
(G)9	-	Hypothetical[13]	S	TP0969 (? & TP0970)
(C)11	?	Hypothetical[14]	S	(TP1030)
2 x (G)8	-	Hypothetical[15]	M	TP0136
(T)9	-	Hypothetical	M	TP0123
(T)9	-	Hypothetical[16]	M	TP0444
(A)9	-	Hypothetical	M	TP0588
(A)9	-	Hypothetical[17]	M	TP0648
(TG)6	-	Hypothetical	M	TP0706
(C)8	-	Hypothetical	M	TP0900
(T)9	+	Hypothetical	M	TP0021 & TP0022
(G)6	(+)	Hypothetical[18]	M	TP0024

+ or - indicated the presence or absence of a frame shift associated with the repeat. pro indicates a repeat located in the putative promoter region of the ORF. S and M indicate strong and moderate candidates respectively.

Notes to Table:

1. The annotated ORF (TP0027) runs from 29838 to 31058. This annotation does not include all of the protein matches and alternative initiation codons are a GTG at 29787, an ATG at 29736 or a GTG immediately 5′ of the repeat. The two genes are transcribed divergently.

2. There is no frame-shift in the annotated ORF but it overlaps the likely start position of the divergent ORF on the other strand. There are 3 alternative start codons in a different frame - the latter of which would accommodate the probable start of the ORF on the other strand. If this is the actual start codon then the repeat is associated with a frame-shift.

3. These genes appear to be present as components of a single transcriptional unit (with a fourth gene located between *mreC* and *pbp-1*). In this context, it is likely that the real start of *mreC* is associated with the ATG located near the termination codon of *mreB* rather than the GTG in the TIGR annotation. This would then associate the repeat with a frame-shift.

4. This ORF has been annotated to start with a GTG which is associated with an overlap between this ORF and the preceding one. There is an ATG alternative 'start' codon that would leave a 1bp gap between this ORF and the termination codon of the one that precedes it.

5. This ORF has been annotated to start with a GTG which results in an overlap with the upstream ORF (TP0144). There is an ATG alternative 'start' codon associated with the termination codon of TP0144.

6. May start at the ATG 13 bp 3′ of the annotated TTG 'start' codon.

7. There is an additional (G)7 repeat in this ORF that is not associated with any frame-shift. Further, there is a termination codon prior to the (G)6 repeat suggesting that this ORF is not functionally intact in the sequenced strain.

8. Annotated with the frame-shift.

9. There is overlap for more than 100 bp between the 3′ end of this frame and the 3′ end of another annotated ORF on the other strand. This other ORF (TP0348) contains a (C)9 repeat, an alteration in the length of which would reduce this overlap.

10. The repeat is located in ORF TP0135. This overlaps with the 5′ located ORF (TP0134) which is likely to be correct because it shared homology with other ORFs in the genome. The repeat is more likely to be associated with an ORF on the other strand which is a homologue of TP0126 and TP0733 that has two frame-shifts and is therefore not functionally intact in the sequenced strain.

11. This is located in the sequence between TP0108 and TP0107.

12. There are ATG 'start' codons in both frames upstream of the repeat which is located 8 bp 5′ of the beginning of the currently annotated ORF.

13. The annotated ORF (TP0969) starts with a TTG. The preceding ORF ends with a termination codon which follows a (T)7 repeat that probably causes a frame-shift in a single ORF composed of TP0969 and TP0970. This ORF is the first of a paralogous family of 4 genes (paralogous family 42) which are arranged in tandem with an organisation that suggests that they are co-transcribed.

14. This represents one of 5 homopolymeric tracts located 5′ of *tpr* genes (*tprC, tprE, tprG, tprJ* and *tprL*). The regions upstream of these genes have some similarities but it is difficult to discern any promoter components. In one instance an ORF has been annotated (TP0318) although this may not represent an expressed protein. The described (C)11 repeat is located in the gene 5′ of *tprL* which is frame-shifted when compared with the *msp* genes described by Fenno *et al.* (1997) at the 5′ end. Extending TP1030 *(tprL)* to include the extent of these genes places the (C)11 in a similar relative location to the repeats 5′ of the other *tpr* genes, although this would substantially overlap the TP1030 sequence on the other strand. It is therefore possible that TP1030 does not represent an expressed protein and thus that the repeat located within it does not mediate phase-variation.

15. These repeats are within a tandem repeat of 2 copies of a 96 mer that is not present in the related ORFs TP0133 and TP0134.

16. The annotated ORF starts with a GTG and includes the (T)9 repeat. However, there is an alternative ATG 'start' 3′ of the repeat that would then no longer be within the translated portion of the ORF and would then be less likely to mediate phase-variation.

17. This ORF (TP0648) is annotated with a TTG 'start' codon that substantially overlaps the 3′ region of the 5′ ORF (TP0649). There is an alternative ATG 'start' 16 bp after the termination codon of TP0649. This does not affect the potential role of the repeat.

18. As annotated TP0024 starts 11 bp after the repeat at a GTG. There is an alternative ATG 'start' codon 48bp upstream of the repeat. It is impossible from sequence analysis alone to determine which represents the true translational start. If the GTG start is correct then this repeat is unlikely to mediate phase-variation.

The interpretation of repeats in an organism for which there is virtually no experimental context is inherently speculative. The inclusion of G or C homopolymeric tracts that are shorter than those composed of A or T is an empirical decision based upon the presence of

ORFs where tracts of this length are clearly associated with a frame-shift. The frame-shift that is observed with a (G)6 in the pyrophosphate-fructose 6-phosphate 1-phosphotransferase (*pfk*) homologue may represent a mutation in the sequenced strain that is unrelated to phase-variation, although there are precedents for instability in repeats of this length (Stibitz *et al.*, 1989; Hammerschmidt *et al.*, 1996b). It is not known to what extent the context of the repeat affects instability or whether the presence of the repeat is the only determinant of instability. It is possible that a repeat of a certain type that is functionally unstable in one location may not be in a different sequence context.

A summary of an analysis of the frequency of homopolymeric tracts in the complete genome is shown in Table 5.2.

Table 5.2: Showing the frequency of homopolymeric tracts in the genome compared with the predicted numbers as determined by Markov-chain analysis and on the basis of the abundance of the previous tract length and the percentage of each base in the genome. The Markov-chain analysis reveals very different results from predictions based upon base composition. For example, the 784 poly-A/T repeats of 7 bases in length is fewer than that predicted by Markov-chain analysis (853) but greater than that predicted on base composition and the frequency of the immediately shorted word / expected from previous (597). It also reveals that the excess of homopolymeric tracts is largely a consequence of the excess of repeats of 7 and 8 bases in length.

Tract length	A or T HPTs	Markov Chain n-1th order	Expected from previous	C or G HPTs	Markov Chain n-1th order	Expected from previous
2	149119		137069	136076		154806
3	51493	41377	36132	34282	30829	35078
4	19263	17781	12473	10003	8636	8837
5	7149	7206	4664	2616	2918	2579
6	2472	2653	1729	625	682	674
7	784	853	597	208	149	161
8	223	248	189	108	68	54
9	44	63	54	62	56	28
10	7	8.2	11	29	34	16
11			1.7	12	13	7.5
12				4	8.7	3.1

The genome of the sequenced strain has an average G + C content of 52.8%. Therefore, if the genome was composed of a random distribution of bases, one would expect that there

would be approximately ¼ as many homopolymeric tracts (or words) for each base of length n+1 than of length n. An alternative (and preferable) approach to the analysis of the frequency of any word is to use a high order Markov chain to estimate the expected frequency of that word. A Markov chain can be thought of as predicting the probability of any word from the occurrence of its component words (*e.g.* the word TGCT contains the component words TGC, GCT, TG, GC and CT). The results of a Markov chain analysis of $n\text{-}1^{th}$ order (n=L-1, where L is length of the word) is shown in table 5.2. The results for the A and T repeats demonstrate the importance of estimating the probability of each word in this way, rather than simply from base frequency on an assumption that sequence is random. For example, there are more AA and TT dinucleotides than CC and GG dinucleotides even though there is a greater than 50% G+C content. Likewise there is an excess of poly-A or -T tri- and tetranucleotides when compared with poly-C or G tri- and tetranucleotides. This will affect the abundance of longer words that include these components.

The analysis of A or T homopolymeric tracts (HPTs) of length 2 to 8 bases, reveals a considerable excess of these relative to their expected frequency estimated from the expected frequency of a HPT of length L, based on the frequency of the HPT of length L-1 with a random distribution of the L^{th} base ('expected frequency from previous' in Table 5.2). However, high order Markov chains reveal that the excess in the tetranucleotides is predominantly a product of the excess of dinucleotides, as shown by the more accurate $n\text{-}1^{th}$ order Markov chain result. Despite an apparent abundance of A and T repeats of 5 to 8 bases in length there are actually fewer than would be expected based upon the frequency of the occurrence of their component words. Whilst an isolated consideration of the frequency of these repeats would suggest a bias towards the generation of HPTs of up to 8 bases, in fact there is a selective pressure against these longer tracts which are, instead, a reflection of the abundance of the di- and trinucleotide repeats.

The analysis of the C or G HPTs of length 2 to 6 bp, reveals a different picture. There are fewer of these than for A or T HTPs, and their frequencies correlate well with the expected frequency from both the Markov chain estimation and the frequency of each previous word. However, there is a marked excess of HPTs longer than 6 bases. These results suggest that there is a bias for the generation of (and hence instability in) C or G HPTs of greater than 6 bp, particularly for C or G HPTs lengths of 7 to 9 in length, longer repeats being approximately according to Markov predictions. Therefore, genes associated with C of G repeats of greater than 6 bp in length may be particularly likely to display phase variability.

The presentation of the results of this analysis has endeavoured to be inclusive but we have attempted to give an indication of what are predicted to be the best candidate phase-variable genes (categorised as strong (S) or moderate (M) candidates). ORFs were included as potentially phase-variable on the basis of the nature and location of the repeat alone and were not influenced by either the TIGR *T. pallidum* gene annotation or the annotation of any potential homologues that were identified by sequence similarity. This approach is appropriate when using a genome in a hypothesis generating exercise such as this, in which one is attempting to identify the nature of contingency genes in a species for which there is little background information.

In other organisms in which repeat-associated phase-variable genes have been investigated they have been predominantly found in association with three broad types of gene: LPS biosynthesis, cell surface proteins and restriction modification systems. *T. pallidum* lacks any identifiable LPS or restriction-modification genes.

T. pallidum has an unusual cell surface that is composed largely of phospholipids with a small number of proteins within it (Radolf *et al.*, 1989 ; Walker *et al.*, 1991 ; Cox *et al.*, 1992). This differs from the other organisms in which repeats have been sought – where the predominant component of the cell surface is LPS. One of the ORFs identified in this analysis, which was not described in the original TIGR annotation (located between ORFs

TP0107 & TP0108), is adjacent to a homologue of *licC* (TP0107), a gene involved in the substitution of LPS with choline in other species, and it may have a similar surface modifying function in *T. pallidum*.

A number of potentially phase-variable DNA metabolism genes other than restriction-modification genes were identified, each associated with a poly-A HPT which may be less prone to slippage than their G or C HPT equivalents. The chromosomal replication initiator protein (*dnaA*) provides recognition specificity for the origin of replication, and is essential for DNA replication (Messer & Weigel, 1996). *T. pallidum* exhibits long latent phases during infection and this may provide a mechanism that produces a sub-population of dormant cells that can act as a source of reactivation. In this context it should be noted that alteration in the length of HPTs does not necessarily require replication. The repeats present in the two error-correction system related genes present another interesting possibility: that the population has the capacity to produce a subpopulation of 'mutator' cells. Increased mutation rates have been identified in human pathogens. This was initially reported in *Esch. coli* and *Salmonella*, where isolates were found to frequently have defective *mutS* genes (Le Clerk *et al.*, 1996). The potential importance of this in host parasite interactions has been recently discussed (Moxon & Thaler, 1997 ; Taddei *et al.*, 1997).

The repeat associated genes with homology with surface proteins that were identified include four flagellar associated genes. The repeat associated with the flagellar motor switch protein (*fliG-1*) is interesting for two reasons. Firstly, it may share its promoter with the divergently expressed hemolysin homologue, reminiscent of the co-phase-varied flagellar genes in *H. influenzae* (van Ham *et al.*, 1993). Secondly, in a similar fashion to *B. burgdorferi* there are two *fliG* gene homologues in *T. pallidum*. The sequences of this pair of genes within *T. pallidum* are substantially divergent which is reflected in the fact that they are annotated on the basis of sequence identity with *fliG* genes from different species. The roles of these two genes, how they might differ functionally, and how altered

expression of one of them might affect motility is not known. However, there may be a similar process of variation in *Borrelia* where the *fliG-2* (BB0290) contains a homopolymeric tract of (A)8 starting at the 8th base of the ascribed reading frame.

In other species the flagellar filament cap protein gene (*fliD*) typically forms a cap structure at the tip of the external filament that is essential for the polymerisation of the filament protein flagellin. FliD-deficient mutants becomes non-motile and leak flagellin monomers into the medium. Phase-variation of this protein might therefore be expected to switch expression of flagellae and hence motility. Further, there are two lines of evidence that suggest other potential roles for FliD in virulence. The first is that it has the potential to function as an adhesin as demonstrated in *Ps. aeruginosa*, in which it is responsible for mucin adhesion (Arora *et al.*, 1998). Since *T. pallidum* has an endoflagellum, FliD is unlikely to be exposed on the surface and to act in a similar way in this species. The second is that *fliD* was detected in a Tn10 mutagenesis study looking for mutants of *S. typhimurium* with impaired survival within macrophages (Baumler *et al.*, 1994). Variation of this gene may be an example of convergent evolution of function with other species. For example, deletion of *fliD* may be the primary cause of loss of flagella in *S. sonnei* (Al-Mamun *et al.*, 1997) and phase-variation of *fliD* (with *fliC*) is responsible for motility variation in *X. nematophilus* (Givauden *et al.*, 1996).

The genes involved in cell wall synthesis appear to be part of a single transcriptional unit (that also includes *mreD* (TP0499) which lies between *mreC* and *pbp1*) and presumably have related functions. This is the first time that this type of gene has been associated with repeats although there are repeat associated potentially phase variable genes of this type in *H. influenzae* (personal observation, see Appendix 1).

The identification of genes involved in general metabolism associated with potentially unstable repeats is also unusual. It may be that genes like *dnaK* (and *dnaJ* for which there is a potential homologue present in the unannotated intergenic sequence which follows *dnaK*) are involved in responses to different environmental conditions. Phase-variable

genes involved in regulation and general metabolism have been identified previously (Stibitz *et al.*, 1989). However, most of these genes are not strong (S) candidates and their variation should be considered only an interesting possibility at this time.

The number of potentially phase-variable genes that do not have identifiable homologues with ascribed functions in the databases is greater than has been found in other organisms. This is probably a reflection of the scale of the difference between the spirochetes and the more intensively studied bacterial species. It is also an indication that *T. pallidum* possesses a number of genes involved in adapting to its host that are novel and the ability to identify candidates from the sequence in this way will hopefully facilitate research in this field.

The issue of surface-associated proteins in *T. pallidum* is controversial. The outer membrane has been found to contain very few protein and the hexameric pore structures that are formed by the major surface proteins (Msp) of other treponemes have not been found in *T. pallidum* (Radolf *et al.*, 1989; Cox *et al.*, 1992). These structural studies have led to a model in which the outer membrane contains few proteins and those that are present are anchored in the cytoplasmic membrane. In this context the finding of repeats in association with glycerophosphoryldiester phosphodiesterase (*glpQ*) is striking. Investigation of the rare outer membrane proteins of *T. pallidum* has identified GlpQ as a surface protein (Shevchenko *et al.*, 1997) and GlpQ has been found to be an immunological target for antibody responses in patients with *B. hermsii* infections, another spirochete (Schwan *et al.*, 1996). Subsequently, a study using opsonic and non-opsonic *T. pallidum* antisera to identify potential opsonic targets also identified this protein suggesting that it is a surface exposed immunological target. Further it has been proposed that it may have immunoglobulin binding capacity in a way similar to its homologue in *H. influenzae* with which it has 72% sequence similarity (Stebeck *et al.*, 1997). ABC transporters have previously been associated with phase variation in *Mycoplasma fermentans* (Theiss & Wise, 1997). These proteins are likely to be membrane associated and might be varied

either as a means of immune evasion or as an adaptation to particular environmental conditions.

The absence of structures typical of Msps in the outer membrane of *T. pallidum* complicates the interpretation of the remaining potentially surface-associated homologues. The *tpr* genes encode the Msp homologues of *T. pallidum* and these have been suggested to act as porins and adhesins on the basis of this homology (Fraser *et al.*, 1998) and are being considered as vaccine candidates (Pennisi, 1998). The occurrence of some of these genes within multigene paralogous families suggests similarities with other families of outer membrane proteins. These include those present in *H. pylori*, and also families of phase-variable genes in other organisms where variable related genes can affect nutrient acquisition or adhesion properties, e.g. iron binding proteins in *H. influenzae*, and the Opa proteins in *N. meningitidis*. A similar approach using the presence of this family of sequences has been used to infer the presence of functional Msp genes in several *Treponema* spp. (Fenno *et al.*, 1997). In apparent conflict with the proposed absence of Msp proteins in the outer membrane of *T. pallidum* there are studies which suggest that, at least during certain stages of infection, they are present and functional. Although antibody alone cannot eradicate organisms during an infection, resistance to reinfection develops during the course of the disease (Magnuson *et al.*, 1956). Monoclonal antibody studies using antibody directed against an Msp-like protein demonstrated that it was reactive in microhemagglutination assays, was capable of blocking attachment of virulent *T. pallidum* to host cells in tissue culture and had 99 – 100% neutralising activity (Jones *et al.*, 1984). These authors concluded that the protein with which the antibody reacted was abundant, immunodominant, surface exposed and important in pathogenesis and as an immunological target. Other studies have also indicated that these proteins are expressed and associated with the cell surface (Marchitto *et al.*, 1984 & 1986; Norgard *et al.*, 1986; Peterson *et al.*, 1986).

It has been suggested that the results that show these proteins to be present on the strains of *T. pallidum* that they have studied are artifactual due to the manner in which the organisms were handled (Cox *et al.*, 1992). This suggestion does address the functional effects of antibodies. An alternative explanation might be related to the phase variation of these proteins and/or proteins associated with their expression and surface location. TP0617/TP0618 and TP0127 are part of a family of genes (paralogous family 14) which also includes two ORFs that do not have repeats suggestive of phase-variation: TP0619 and a duplication of TP0619 that has a frame-shift due to a deletion (TP0314/TP0315). These genes are present near to, and are apparently transcriptionally associated with, the *tpr* genes - which are the family of genes (*tprA* to *tprL*) that encode the treponemal major surface proteins. This situation may be further complicated by TP0479 and TP0697 (paralogous family 30) which are both potentially phase-variable. TP0697 has homology with TP0127 and TP0346. As described above, TP0127 is part of paralogous family 14. TP0346 is part of paralogous family 21 with TP0347, another potentially phase-variable gene. So there may be a larger family of genes, with functions that may be associated with those of the *tpr* genes, of which several members: TP0127, TP0618/0619, TP0347, TP0479 and TP0697 are strong candidates for phase-variation. If the expression of the Msp proteins were phase variable this might explain their absence in some studies and their presence in others. According to this model, only a small subpopulation of bacteria may be expressing these proteins within a population under conditions when they are not required or when immune responses are directed against them and their expression may only be favoured under particular environmental conditions or stages of the infection process.

The ORFs on the reverse strand of TP0135 and TP0126 also appear to represent a pair of related phase-variable genes. However, in several cases the potentially variable genes are members of paralogous families where the other members do not have repeats suggestive of phase-variation including: TP0961, TP0136, TP0026, TP0859/TP0860 (although one of the paralogs: TP0865, contains (G)6 and (A)8 repeats), TP0969, TP0706, TP0024.

This analysis highlights a group of 42 genes (of which 22 are strong candidates) that can be focussed upon in the study of the virulence of *T. pallidum*. The length and nature of the repeats would allow detection of the presence or absence of variation in the length of repeat tracts directly from organisms obtained from different disease sites and from models of infection by PCR and sequencing. This provides a novel approach to the study of this organism that does not require a culture system as a pre-requisite. It also demonstrates the use of a complete genome sequence to provide new starting points for the investigation of the biology of a pathogenic bacterium.

Chapter 6

Repeat associated phase variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58.

6.1 Introduction

Pathogenic *Neisseria* species, *N. meningitidis* and the closely related *N. gonorrhoeae*, are responsible for causing bacterial meningitis and gonorrhoea respectively and remain important human pathogens despite the availability of antibiotics to which they are susceptible. One of the major barriers to their control is the inability to develop effective vaccines against them. The structures expressed on the cell surface of the pathogenic *Neisseria* species have been extensively studied and the majority display substantial intra-strain variation. Such diversity provides a capacity to adapt to different niches both within individual and different hosts, to escape immune responses and to influence the various stages of the bacterium-host interaction. Phase variation is arguably the most important mechanism influencing intra-strain diversification in these species.

Phase variation describes a process of reversible phenotypic switching that is mediated by DNA alterations or modifications, characteristic of host interactive determinants, referred to as contingency genes, that adapt the bacteria to different microenvironmental conditions (Moxon *et al.*, 1994). In meningococci, one of the phase varied bacterial components is the capsule. When the capsule is expressed it confers serum resistance, while loss of expression facilitates adhesion to cell surfaces mediated by surface proteins (DeVoe, 1982; Virji *et al.*, 1992a & 1993a, Stephens *et al.*, 1993; Hammerschmidt *et al.*, 1994 & 1996a). Pili, which project beyond the capsule and make initial contact with epithelial cells, are required for adhesion of encapsulated bacteria to epithelial and endothelial cells (Stephens & McGee, 1981; Virji *et al.*, 1991; Nassif *et al.*, 1994). Several features of pilus biogenesis and function are phase variable, including their expression, associated proteins mediating

adhesion, and substitutions of the pilus subunit. The latter feature is mediated both through changes in the pilin-subunit substrate and the expression of modification enzymes (Rudel *et al.*, 1992 & 1995a; Virji *et al.*, 1993b; Nassif *et al.*, 1994; Weiser *et al.*, 1998b; Jennings *et al.*, 1998). Further cell adhesion and tropism are also determined by other surface proteins including the Class 5 proteins (Opa and Opc proteins) (Virji *et al.*, 1992a & 1993a). The Opa proteins are a family of divergent adhesion proteins (four present in *N. meningitidis* strain MC58) that, in addition to being independently phase variable, are also capable of variation through recombination to generate novel Opa variants (Sparling *et al.*, 1986; Stern *et al.*, 1984 & 1986; Achtman *et al.*, 1988; Bhat *et al.*, 1991). The nature of host cell interactions differs depending upon which variant Opas are expressed (Waldbeser *et al.*, 1994; Kupsch *et al.*, 1993; Virji *et al.*, 1993a; McNeil *et al.*, 1994). Opc is also phase variable and, in those strains in which it is present, has been shown to be the most important determinant of cellular adhesion and invasion studied to date, promoting adhesion to epithelial cells as effectively as the most adherent Opa protein (Virji *et al.*, 1992a & 1993a; Sarkari *et al.*, 1994). The Class I protein (PorA), a pore forming protein with cationic selectivity, is the most abundant cell surface protein. It also shows variation in expression in clinical isolates and is phase variable (Tommassen *et al.*, 1990; Poolman *et al.*, 1980; Hopman *et al.*, 1994; van der Ende *et al.*, 1995). The acquisition of iron, present in low concentrations in the host, is essential and involves several systems in pathogenic *Neisseria* (Schryvers & Stojiljkovic, 1999). Several of the Neisserial iron binding proteins have been found to be phase variable (Chen *et al.*, 1996 & 1998; Lewis *et al.*, 1999).

In addition to capsule, pili, surface proteins, and surface protein modifications, the lipopolysaccharide (LPS), which is the predominant component of the outer membrane, also displays phase variation. Initially studied in *N. gonorrhoeae*, LPS was found to be size and antigenically heterogeneous in a single strain and to vary during experimental infection (Apicella *et al.*, 1987; Schneider *et al.*, 1988 & 1991; Weel *et al.*, 1989). Several phase

variable LPS biosynthetic genes have now been identified in *Neisseria* and these are capable of generating a wide variety of LPS phenotypes (van Putten & Robertson, 1995; Jennings *et al.*, 1999). In meningococci there is an association between LPS phenotype and invasive disease, which is at least in part related to the presence of a substrate for sialylation of the LPS dependent upon the expression of *lgtA* (Jones *et al.*, 1992; van Putten 1993; Danaher *et al.*, 1995). These changes in LPS structure alter interactions with host cells and can also mask some LPS and protein epitopes (Poolman *et al.*, 1988; Judd & Shafer, 1989; van Putten, 1993; de la Paz *et al.*, 1995).

The contribution of phase variation to the expression of the major surface components of pathogenic *Neisseria* influencing host interactions and hence virulence is well established. In addition, a recent study has suggested that differences in the presence of the Dam methylase involved in mismatch repair affects the rates of phase variation and the loss of Dam is associated with virulent invasive disease isolates (Bucci *et al.*, 1999). The relative fitness advantages in different micro-environments of altered phenotype through expression, and immune evasion through the absence of expression is unknown. However, in each case the existence of variability indicates the central role of the varied structure in the host-bacterial interaction. Since more phase variable genes are recognised in the pathogenic *Neisseria* than in any other species, it seems possible that most if not all of these would already be characterised in these intensively studied organisms. Less abundant surface proteins than those which have been intensively studied may also be immunological targets on the cell surface (Manning *et al.*, 1998). In addition, a study of the contribution of phase variation to invasion of primary nasopharyngeal cell lines suggests the presence of additional uncharacterised phase variable proteins involved in this process (de Vries *et al.*, 1996). The availability of a complete genome sequence makes a comprehensive investigation of this type possible (Saunders & Moxon, 1998).

We have analysed the complete genome sequence of the virulent serogroup B *N. meningitidis* strain MC58 (which is Dam-negative) in order to fully characterise the

repertoire of repeat associated phase variable genes in this virulent pathogen and to identify putative novel determinants of host-bacterial interactions.

6.2 Methods

The complete genome sequence of *N. meningitidis* strain MC58 (Tettelin *et al.*, 2000) was processed and interrogated using the previously described genome analysis software (Saunders *et al.*, 1998). Simplicity within the genome was surveyed using a number of overlapping methods: **ARRAY** (a program that searches for all perfect repeats with 5 or more copies of the component motif, in which motifs are less than 10 bases in length), **SIMPLEP** (a program that uses a probabilistic method to identify sections of genome that are more repetitive than would be expected by chance), **TANDEM** (used to identify 10 to 250 base imperfect direct repeats) and **QUICKTANDEM** (used to identify 250 to 1000 base direct repeats). The whole genome was searched for sequence identity using **BLASTN, BLASTX** and **TBLASTX** against archeal, prokaryotic and lower eukaryotic sequences (i.e. excluding metazoa). Matches were divided into 'self' (*Neisseria* spp.) and 'non-self' (non-*Neisseria* spp.). The results of these analyses were loaded into ACEDB, interrogated by two independent operators, and the results were combined. These results were cross-checked with the tabulated results from the ARRAY and TANDEM analyses.

In addition, the frequency of each homopolymeric tract up to 12 bases in length was used to determine the expected numbers of each homopolymeric tract length using high order Markov chains (Cox & Miller, 1965) to determine the expected frequency of sequence 'words' based upon the frequency of their component parts. In this way 'words' can be determined to be present at higher or lower frequencies than expected (Saunders *et al.*, 1999b). Based upon the Markov chain analysis and known unstable repeats in *Neisseria*, repeat sequences were selected if they consisted of homopolymeric tracts of greater than 6 Gs or Cs, 8 As or Ts, 4 copies of dinucleotides (5 for GC/CG), 3 copies of tetramer and longer motifs, and all repeats of 5 or more bases associated with frame shift changes.

These repeats were interpreted on the basis of sequence context and the effects of altered repeat length on associated reading frame expression.

6.3 Results and Discussion

The results of the Markov-chain type analysis of homopolymeric tracts are presented in Figure 6. This reveals that there are fewer than expected G/C homopolymeric tracts of less than 5 bp and a large excess of homopolymeric tracts greater than 6 bp, and particularly 8 bp, in length. Note that for each repeat the excess of the components parts is accounted for in the prediction. This suggests that there is a strong mechanistic bias for the generation and instability of these repeats. In contrast, there is an abundance of A/T trinucleotides (and dinucleotides – data not shown) but longer repeats of up to 8 bp are less frequent than predicted suggesting that they are selected against, whilst there is a moderate excess of longer repeats of this type.

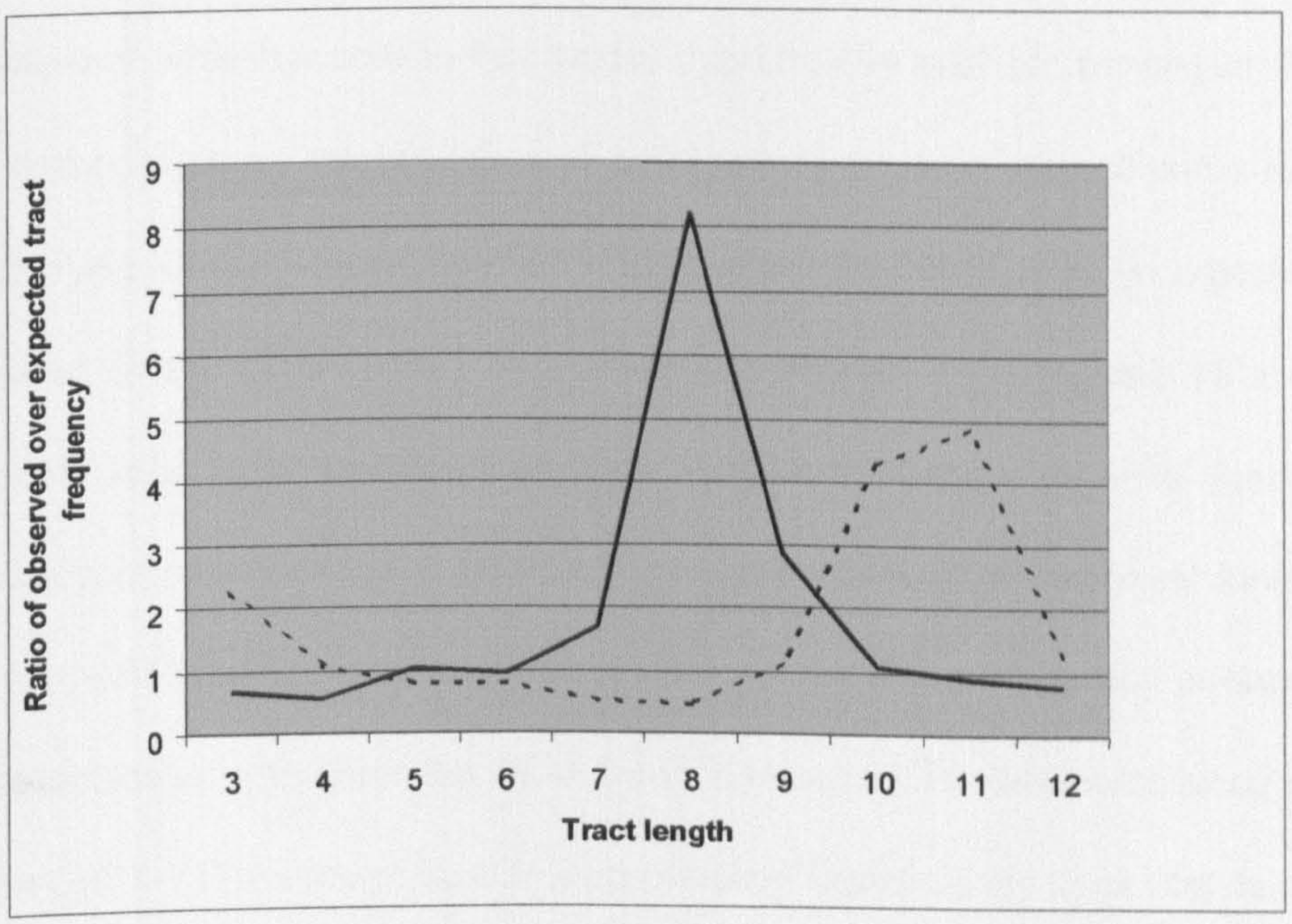
Repeats present in a sequence context expected to affect the expression of the associated open reading frames are listed in Table 6. The most striking finding is the number of putative phase variable genes present in this genome sequence. There are 66 potentially phase variable genes identified in this analysis of which 13 are previously proven to be contingency genes, 31 are newly identified strong candidates, and there are an additional 22 possible phase variable genes. This contrasts with a total of 15 potential phase variable genes in *H. influenzae*, 26 in *H. pylori*, and 42 (of which 22 are strong candidates) in *T. pallidum* (Hood *et al.*, 1996; van Belkum *et al.*, 1997a; Saunders *et al.*, 1998; Saunders – unpublished).

One notable feature of the analysis of *N. meningitidis* strain MC58 is that a large proportion of the phase variable genes with repeats within the open reading frame (53%; 30 of 57) have repeat lengths that would not result in expression. In part this is possibly a reflection of *in vitro* cultivation and the absence of selective pressures which favour expression. In the context of this analysis this actually provides some support for the phase

Figure 6.

Markov chain – type analysis of homopolymeric tracts in *N. meningitidis* strain MC58.

Figure showing the ratio of observed and expected homopolymeric tract frequencies using word length L-2 for the predictions (Gs and Cs – solid line; As and Ts - dashed line).



variability of these genes and the instability of the particular types and lengths of repeat observed. The identification of phase variable contingency genes and predictions of repeat instability made on the basis of this type of analysis in other species have been supported by experimental and comparative studies (Hood *et al.*, 1996; van Belkum *et al.*, 1997a & b; Alm *et al.*, 1999). These previous studies proceeded with less contextual information with which to design the search parameters and to interpret the results. In contrast, for *N. meningitidis* there is already a substantial body of knowledge on the length, composition and instability of particular repeats known to be associated with phase variation. The composition of repeats previously associated with phase variation in *N. meningitidis* is consistent with that seen in this study. Functionally unstable repeats tend to be composed of homopolymeric tracts of G or C and longer repeats which include a mixture of purines and pyrimidines. The exception to this pattern are the (CA/TG)n repeats including those located in the OMP85/D15 and restriction enzyme homologues. This is different from repeats seen in *Neisseria* or in other species in association with functional instability. Sequences of OMP85/D15 from *N. meningitidis* and *N. gonorrhoeae* have been previously published (Manning *et al.*, 1998) but these do not demonstrate any polymorphism between themselves or with strain MC58 at the (CA)4 repeat. In addition most of the repeats of this type ((CA/TG)4) are present in open reading frames in contexts that do not suggest phase variation and/or in housekeeping genes that are unlikely to be variable. The genes with this type of repeat have been included in this analysis on the basis of the repeat location and length but they should only be considered to be moderate candidates until supported by experimental data. In contrast, (AT)n repeats have been associated with phase variation in *H. influenzae* (van Ham *et al.*, 1993) and the (AT)5 repeat in the *fixP* homologue is one of only 2 repeats of this type/length in the genome and the only one that is located within an open reading frame. Furthermore, even though the majority of the sequence is coding, half of the (AT/TA)4or5 repeats are located intergenically which supports the hypothesis that this sequence is selected against in open reading frames, possibly due to instability.

Although functional variability exerted through changes in poly-A/T tracts has been reported in *Mycoplasma* spp. (Theiss & Wise, 1997; Zhang & Wise, 1997) it has not been documented in *Neisseria* spp.. Therefore, genes associated with repeats of fewer than 9 As or Ts have not been considered to be strong candidates for phase variation. The case for the longer poly-A/T repeats is enhanced by the presence of a repeat-associated frame shift at the (A)8 in *mesJ*, and that repeats of this type are predominantly located in intergenic regions. Their abundance also suggests that they do not alter in length at high rates.

Phase variation is a characteristic of the LPS of *Neisseria* and is thought to play a role in virulence, as variation in LPS phenotypes may help adapt cells to particular niches or facilitate evasion by host mimicry (Giardina *et al.*, 1999). The observation of previously unrecognised glycosyltransferases in the genome sequence suggests the presence of novel LPS structural variants and/or other saccharide modifications of other surface features. The exceptionally long heptameric repeat associated with one of these genes is the longest repeated motif identified to date and further extends the combinatorial potential of the variable LPS genes. Not every LPS gene recognised in other strains to be phase variable has repeats that are long enough to mediate high frequency variation in strain MC58. The *lgtB* and *lgtE* genes do not have sufficiently long repeats to be unstable at a high frequency in strain MC58, having homopolymeric tracts of only 5 bases in length. The relative stability of these genes when they have repeats of this length has been demonstrated experimentally (Jennings *et al.*, 1999). Different strains have different variable repertoires of LPS phenotypes on the basis of variations in the genes present and those which have long repeat tracts. This pattern is consistent with a situation in which different strains vary in their repertoire of potential LPS phenotypes through the variable presence of alleles and also in their repertoire of diversification through strain to strain differences in which genes are phase variable. It is notable that the strain MC58 *lgtB* and *lgtE* genes do have (G)5 repeats at the locations in which longer tracts have been reported. These repeats are relatively unusual in coding sequence suggesting that these are remnants of longer repeats

from an ancestral strain that was able to vary these structures. It is possible that the presence of these repeats may act as a 'genetic memory' that predisposes the genes to regain phase variable potential under appropriate selective conditions. A similar situation may apply to other genes such as the acetyltransferase (NMB 0285/6) where slippage in the (G)6 repeat may restore function infrequently.

A similar pattern of differences in gene repertoire and phase variability emerges when the iron binding proteins are considered. The hemoglobin binding protein *hmbR* has recently been recognised to be phase variable (Lewis *et al.*, 1999; Richardson *et al.*, 1999). *frpB* is an iron regulated gene involved in the utilisation of ferric ions from a number of sources including transferrin and lactoferrin. Expression of FrpB was recognised to be phase variable due to a promoter located repeat in *N. gonorrhoeae* but not previously in *N. meningitidis* for which the previously published sequence starts 3′ of the repeat location (Dyer *et al.*, 1988; Beucher & Sparling, 1995; Pettersson *et al.*, 1995). The phase variability of this gene in *N. meningitidis* may affect its candidacy for vaccine development (Ala'Aldeen *et al.*, 1994; van der Ley *et al.*, 1996). The lactoferrin binding protein gene (*lbpA*) has been sequenced previously from *N. gonorrhoeae* but the sequence has only 5Gs at the location of the 8Gs present in *N. meningitidis* strain MC58. This gene has not been previously considered to be phase variable by this mechanism (Biswas & Sparling, 1995) and probably represents a situation similar to that of the *lgt* genes in which some strains have reduced tract lengths. Indeed a previously published meningococcal *lbpA* does not have any repeat at all – the longest run of Gs being only 2 bases in length (Pettersson *et al.*, 1993). This represents the most unequivocal example of a gene that has features of phase variability in one strain and stability in another. Finally, the phase variable gene involved in hemoglobin, haemoglobin-haptoglobin and apo-haptoglobin utilisation (*hpuA*) previously described in *N. gonorrhoeae*.and *N. meningitidis* strain DNM2 is absent in *N. meningitidis* strain MC58 (Chen *et al.*, 1998; Lewis *et al.*, 1999). *N. meningitidis* strain MC58 has several iron acquisition systems (Tettelin *et al.*, 2000) and it is likely that there

is functional redundancy within and between different strains of *Neisseria*. The presence of different genes related to common functions in different strains, and differences in the phase variable repertoire of these genes, highlights the enormous diversity that exists between strains of the same species.

The identification of additional novel surface proteins with homology to phase variable virulence determinants from other species is interesting. In particular it is notable that variability of the *yadA/yopI* homologue may represent convergent evolution with the homologous gene in *H. influenzae* which is also associated with repeats (Hood *et al.*, 1996). It is not possible to predict what proportion of the open reading frames for which there are no interpretable homologies will also prove to encode surface associated proteins. These potentially include a substantial proportion of the greater than 20 immunogenic minor surface proteins reported to date (Manning *et al.*, 1998).

Phase variable restriction modification systems have not been previously reported in *N. meningitidis*. There is a single report of a repeat associated type III restriction/modification enzyme in *N. gonorrhoeae* (Belland *et al.*, 1996). Phase variable restriction/modification enzymes have been identified previously in other species including *Mycoplasma*, *H. influenzae* and *H. pylori* (Dybvig & Yu, 1994; Hood *et al.*, 1996; Saunders *et al.*, 1998) although a variable specificity protein gene has not been identified before. With the exception of the phase variable type I system in *Mycoplasma bovis* that is associated with variable resistance to bacteriophage, which is switched by a different mechanism, the function of these variable restriction-modification systems is unknown (Dybvig & Yu, 1994). On the basis of this association with 'phage resistance and susceptibility these genes have been categorised, with the 'phage and bacteriocin associated genes, as 'Bacterial population competition determinants', but they may have other or additional functions. The number of 'moderate' candidates is also worthy of note. Other than the tetramer and pentamer associated enzymes from this analysis, there is a trend towards restriction systems having shorter repeats than are seen in other genes. These may have a lower

frequency of length variation and reflect a selective advantage associated with variation at lower rates than are typical of other phase variable proteins. Even more striking are the candidate bacteriophage and bacteriocin related proteins which have not previously been associated with phase variation. The *funZ* gene is a homologue of a gene that is involved in lysogenic conversion in 'phage P2 [GI3139107] whilst the Ner protein from 'phage Mu is a DNA binding protein which may function as a repressor protein (Strzelecka *et al.*, 1995). Variation in the expression of these genes would be expected to affect the susceptibility to and productivity of bacteriophage infections thus producing populations of bacteria in which the 'phage replicates and other populations in which it is harboured silently. The functional consequences of phase variation of both restriction systems and bacteriophage proteins may overlap, generating mixed populations of resistant and susceptible populations and latent and lysogenic infections respectively, each type of variable phenotype being adaptive to the other. Phase variable bacteriocin expression would potentially generate sub-populations that are intermittently benign and lytic to others. This is particularly interesting in the context of a naturally transformable species that can incorporate DNA from other cells from the same and different strains. Taken together this repertoire of candidate phase variable genes extends the role of phase variation beyond the recognised area of host-bacterium interactions and extends it to include variability in the relative fitness of sub-clones within bacterial populations.

There are more novelties in the types of gene identified in this analysis. Secreted enzymes, toxins and toxin secretion systems have not been recognised to be potentially phase variable virulence determinants in *Neisseria* spp. and have only previously been described in *Bordetella pertussis* (Gross & Rappuoli, 1989). The variation of the di-heme cytochrome C would potentially adapt the population to different microenvironmental conditions in a novel way. This protein is an essential component of a terminal receptor of an apparently branched electron transport chain that would be expected to adapt the organism to microaerophilic conditions, similar to that of the endosymbiotic bacteroids in

rhyzobial species where its oxygen affinity allows respiration at low oxygen concentrations (Preisig *et al.*, 1993 & 1996; Thony-Meyer *et al.*, 1994; Koch *et al.*, 1998). This particular type of oxidase has been reported to have a less tight coupling between oxygen reduction and proton translocation (de Gier *et al.*, 1996). Therefore, the expression of this gene may be metabolically disadvantageous when the bacterium is not in the environment for which this complex is adaptive. The adaptive benefits of variation in *nifS* which affects Fe-S complex formation, a glutaredoxin which reduces di-sulphide bonds, and a phospholipid biosynthesis gene that might affect cell envelope biosynthesis are currently unclear.

The number and functional diversity of known and candidate phase variable genes in *N. meningitidis* is unparalleled in any species investigated to date. Each bacterial clone is able to potentially explore many thousands of phenotypes due to the independent switching and hence combinatorial nature of this process. The results of this analysis suggest that the role of phase variation in bacterial fitness can be extended beyond that of adaptation to surface conditions and evasion of host responses to include competitive fitness within the bacterial population and adaptation to metabolic microenvironmental conditions. This extension does not alter the model of phase variation as it relates to the function of contingency genes (Moxon *et al.*, 1994). Whilst this analysis identifies many new contingency genes, it also increases our understanding both of the phenotypic flexibility of *N. meningitidis* and the role of stochastic switching processes in adaptation in general within bacterial populations to local environmental conditions.

Table 6.

Repeat associated putative phase variable genes in *N. meningitidis* strain MC58.

Repeat	Frame shift	Gene similarities		NMB number
Surface associated proteins				
(G)11	+	Pilus assembly protein (*pilC2*)	K	NMB0049
(G)14	+	Pilus assembly and adhesion protein (*pilC1*)	K	NMB1847
(G)11	P	Class I outer membrane protein (*porA*)	K	NMB1429
(C)12	P	Class 5 protein / surface adhesion protein (*opc*)	K	NMB1053
(CTTCT)10	+	Class 5 protein / surface adhesion protein (*opa*)	K	NMB0442
(CTTCT)11	+	Class 5 protein / surface adhesion protein (*opa*)	K	NMB1636
(CTTCT)13	+	Class 5 protein / surface adhesion protein (*opa*)	K	NMB1465
(TCTTC)16	+	Class 5 protein / surface adhesion protein (*opa*)	K	NMB0926
(G)6	+	'Cell adhesion molecule' – from patent match	S	NMB2104
(C)9	-	Outer membrane protein related to adhesion / invasion proteins and IgA protease	S	NMB1998
(TAAA)9	P	Outer membrane protein (*yop1/yadA* related)	M	NMB1994
(GGCA)3	-	Adhesion and penetration protein homologue	M	NMB1985
(AC)4	-	Outer membrane protein homologous to D15 (*omp85*)	M	NMB0182
(AC)4	-	Transporter	M	NMB1277
(G)9	-	Hemoglobin receptor (*hmbR*)	K	NMB1668
(G)8	-	Lactoferrin binding protein (*lbpA*)	S	NMB1540
(C)11	P	Iron acquisition protein (*frpB*)	S	NMB1988
Surface sugar biosynthesis proteins				
(G)13	-	Saccharide acetylase	S	NMB1836
(CAAACAA)34	+	Glycosyltransferase	S	NMB0624
(G)14	-	LPS glycosyltransferase (*lgtA*)	K	NMB1929
(C)12	+	LPS glycosyltransferase (*lgtG*)	K	NMB2032
(C)7	-	Capsule biosynthetic protein (*siaD*)	K	NMB0067
(G)11	-	Pilus glycosyltransferase (*pglA*)	K	NMB0218
Toxin and secreted enzyme related				
(ATAACAAA)4	+	RTX-type toxin*	S	NMB1407
(C)10	-	Serine protease	S	NMB1969
Bacterial population competition determinants				
(G)7	+	Restriction modification system specificity protein (*hsdS*)*	S	NMB0831
(A)9	+	Type I restriction modification system modification protein (*hsdM*)*	M	NMB1223
(TG)4	-	Type II restriction enzyme	M	NMB0726
(G)6	-	Type II restriction enzyme	M	NMB1032
(CAGC)20	+	Restriction-modification system modification protein (*mod*)	S	NMB1375
(CCCAA)16	+	Type III restriction modification system modification protein (*mod*)	S	NMB1261
(CAAAT)5	-	Bacteriophage gene (*funZ*)	S	NMB0961
(A)7	+	Bacteriophage protein (*ner*)	S	NMB1080

(G)7	+	Bacteriocin export protein (*mtfB*)	S	NMB0098
(C)5	+	Colicin V secretion protein (*cvaA*)	S	NMB1783
Others				
(AT)5	-	Di-heme cytochrome C (*fixP*)	S	NMB1723
(C)8	-	Protein involved in Fe-S complex generation (*nifS*)	S	NMB1379
(TGCG)3	-	Glutaredoxin 2	M	NMB1734
(TTCC)3	-	Fatty acid / phospholipid sythesis protein (*plsX*)	M	NMB1913
(A)8	+	Cell cycle protein (*mesJ*)	M	NMB1140
Proteins of unknown function and hypothetical proteins				
(AAGC)9	+	FUN (*nmrep2*)	R	NMB0312
(AAGC)5	+	FUN (*nmrep3*)	R	NMB1525
(G)9	+	FUN	S	NMB0415
(G)7	+	FUN	S	NMB0486
(C)7	+	FUN	S	NMB0970
(G)7	+	FUN	S	NMB1893
(TTCC)4	(+)	FUN	S	NMB1741
(G)7	+	FUN	S	NMB0593
(C)8(N9)(G)7	P	FUN	M	NMB1634
(C)8(N)10(G)7	P	FUN	M	NMB1543
(AT)4	-	FUN (transmembrane protein)	M	NMB0432
(GAAA)3	-	FUN	M	NMB1265
(AC)4	-	FUN	M	NMB0471
(CAAG)11	-	Hypothetical protein (*nmrep1*)	R	NMB1507
(AGCA)3	(+)	Hypothetical protein	S	NMB1275
(C)7	-	Hypothetical protein	S	NMB0488
(C)7	-	Hypothetical protein	S	NMB1489
(G)7	(+)	Hypothetical protein	S	NMB1931
(G)6	(+)	Hypothetical protein	S	NMB0300
(C)6	(+)	Hypothetical protein	S	NMB1760
(A)11	-	Hypothetical protein	S	NMB0368
(T)10	-	Hypothetical protein	S	NMB0065
(C)7	P	Hypothetical protein	M	NMB2008
(A)9	P	Hypothetical protein	M	NMB1786
(A)11	P	Hypothetical protein	M	NMB0032

Notes: * open reading frame inactivated by other mutations. + and - indicate the presence or absence of a frame shift in the respective ORF, (+) indicates a frame-shift in the most probable ORF in cases where this cannot be implied from homologies, P indicates promoter located repeat. K = previously recognised phase variable gene; S = strong candidate; M = moderate candidate; R = previously documented repeat for which the associated gene has not been confirmed to be phase variable.

Chapter 7

Investigation of the mechanism of phase variation of *opc* in serogroup B *N. meningitidis*.

7.1 Introduction

Opc is one of the class 5 proteins of *Neisseria meningitidis*. Class 5 proteins are abundant, basic, trimeric, surface proteins that share the characteristics of being heat modifiable and frequently changing their expression, independently, in a phase variable fashion (Tsai *et al.*, 1981; Poolman *et al.*, 1980; Frasch & Mocca, 1978; Crowe *et al.*, 1989; Woods & Cannon, 1990). Opc, originally named protein 5C, is expressed by many meningococci belonging to different serogroups and differs from the other Class 5 proteins in that it has a constant subunit size and antigenicity. In contrast, Opa proteins undergo recombination to generate new variant Opa protein subunits (Connell *et al.*, 1988). Opc also differs from the Opa proteins in the nature of its phase variation. Opa proteins exhibit ON-OFF switching with no intermediate phenotypes, a process mediated through changes in the number of pentameric repeats in the open reading frame (Stern *et al.*, 1986). Opc differs in that bacteria can express intermediate Opc phenotypes (Achtman *et al.*, 1988; Virji *et al.*, 1992a). Bacteria expressing large amounts of Opc were preferentially isolated from the nasopharynx whilst systemic isolates preferentially expressed small amounts of the protein, which has been interpreted to suggest that Opc plays a role in initial adhesion events during colonisation (Achtman *et al.*, 1991a). *In vitro* studies support this model and demonstrate that Opc can mediate adhesion to, and invasion of, endothelial and epithelial cells (Virji *et al.*, 1992a & 1993a). Opc is as effective at mediating interactions with epithelial cells as any of the Opa proteins with which it has been compared and it is significantly more efficient than Opa proteins in mediating interactions with endothelial cells. The interaction of Opc expressing meningococci with endothelial cells has two

patterns which suggests at least two mechanisms by which Opc can mediate bacterium – host cell adhesion. There is a polar adhesive pattern that requires an intermediate ligand and involves surface integrins, and there is a second pattern that is non-polar that is not influenced by factors which specifically inhibit the integrin mediated adhesion. Vitronectin can function as the intermediate ligand and antibodies raised against the αvß3 family of integrin receptors are the most efficient antagonists of the polar Opc mediated binding (Virji *et al.*, 1994). Opc also binds directly to cell surface proteoglycans, which includes cell surface associated heparin (de Vries *et al.*, 1998). This may be the mechanism of non-polar binding. The combination of these specificities may be functionally important because studies of OpaA expressing *N. gonorrhoeae* suggest that host cell proteoglycans are needed for bacterial adherence, but that vitronectin is required to complete bacterial entry into Chang epithelial cells (Duensing & van Putten, 1997). As revealed by adhesion patterns which differ from those described above, there may be at least one more mechanism by which Opc mediates adhesion. During experiments in which high density polar binding to endothelial cells was inhibited by RGD peptides, which antagonise binding to integrins, there was a residual sub-population of cells which still bound large numbers of bacteria (Virji *et al.*, 1994). Since *N. meningitidis* is associated with invasive disease and septicaemia, a process that has to involve crossing endothelial cell layers, it is tempting to see the documented interactions with endothelial cells in this context. However, it is probably more appropriate to consider the endothelial cells more simply as model cells which happen to express receptors with which Opc can directly and indirectly interact. The process of invasion to the lumen of a vessel involves migration from the basolateral to the apical surface, which is the reverse of the process modelled in endothelial cell culture. Further, Opc is predominantly expressed by nasopharyngeal isolates rather than those from deep sites, and this is the reverse of what would be expected if Opc had a central function in migration to and from the vascular compartment.

Opc, although poorly immunogenic in experimental animals, is highly immunogenic in humans and elicits bactericidal antibodies (Achtman *et al.*, 1988; de Cossio *et al.*, 1992; Rosenqvist *et al.*, 1993 & 1995). Opc also contains several strong T-cell epitopes in the transmembrane domains (Wiertz *et al.*, 1996). High level Opc expression in strains used to prepare intranasally administered outer membrane vesicle vaccines was associated with greatly increased bactericidal responses when compared to weakly expressing strains, even in the absence of measurable increases in serum IgG (Haneberg *et al.*, 1998). However, the level of Opc expression of the bacteria used in the bactericidal assays in this study was not specified, making this difficult to interpret. In addition to avoidance of antibody responses, phase variation may also be important in avoiding interactions with cell mediated immunity. In contrast to Opa proteins, the presence of Opc does not elicit responses from polymorphonuclear phagocytes (McNeil & Virji, 1997). However, expression of Opc leads to rapid uptake and killing by monocytes which is dramatically reduced in cells which no longer express Opc (McNeil *et al.*, 1994).

opc was cloned from an Opc positive revertant of an isolate from the Gambia that previously expressed no class 5 proteins (Crowe *et al.*, 1989) by generating a lambda library and selecting an *Esch. coli* transformant that expressed Opc efficiently using anti-Opc monoclonal antibodies (Olyhoek *et al.*, 1991). The resultant plasmid was called pBE501, and the antibody used was called, A222/5b. The expressed Opc protein was confirmed to be surface located in *Esch. coli* by immunofluorescence. Sequencing of pBE501 revealed an ORF encoding a protein of 269 amino acids, with a 19 amino acid leader peptide typical for membrane proteins, and with a predicted final mass of 28kDa. *opc* was found to have less than 35% DNA homology with *opa*s and to lack the pentameric repeats known to mediate phase variation of the Opa proteins. It was concluded that these proteins are derived from a single protein family but have diverged considerably. The deduced Opc amino-acid sequence was used to devise a two-dimensional structural model of Opc which contains 10 transmembrane strands and 5 surface exposed loops (Merker *et*

al., 1997), which contrasts with the Opa proteins which are thought to have only 4 surface exposed loops (van der Ley, 1988; Bhat *et al.*, 1991). The surface availability and functional roles of loops 2, 4 and 5 were confirmed by binding to monoclonal antibodies that were both bactericidal and inhibited adhesion and invasion. Subsequently a projection structure has been determined for Opc in lipid vesicles which is consistent with a trimeric pore forming structure with extensive surface projections as suggested by the earlier studies (Collins *et al.*, 1999).

The expression of Opc is phase variable in the laboratory and in nature (Crowe *et al.*, 1989; Achtman *et al.*, 1991a; Virji *et al.*, 1993a). As stated above, phase variation of Opc differs from that of the other class 5 (Opa) proteins in that the amount of Opc protein expressed varies quantitatively (Achtman *et al.*, 1988; Virji *et al.*, 1992a). This has been ascribed to variation in the length of a homopolymeric tract of C residues located 5′ of the –10 region of the *opc* promoter (Sarkari *et al.*, 1994). Opc expression was found to correlate with three levels of transcription: strong expression; intermediate/weak expression, with a 5 to 10 fold reduction of transcript compared to strong expression; and no detectable mRNA in the Opc-OFF variants. Sequencing of different isolates having different Opc expression showed a correlation between Opc expression and the number of Cs in the homopolymeric tract. In this study of 65, predominantly serogroup A, strains the majority conformed to an association of 12 and 13 Cs with strong expression, 11 and 14 Cs with weak expression, and <11 or >14 Cs with the Opc-OFF phenotype. There were 3 strains within the studied collection that did not conform to this model. The figure in the paper which accompanies this description shows that the number of Cs on the sequencing gels is difficult to read and becomes progressively more indistinct with the length of the homopolymeric tract that is sequenced. Finally, variation between strong and weak expression was shown to be associated with changes in the homopolymeric tract between 13 and 14 Cs. It is noteworthy that in this study strain C751, from which pBE501 was generated in the previous cloning and expression experiment in *Esch. coli* was found to

have 14Cs, whilst pBE501 which was associated with expression in *Esch. coli* had only 10 Cs in the homopolymeric tract. According to this observation the homopolymeric tract length associated with expression in *Esch. coli* and *N. meningitidis* would appear to differ.

The presence of intermediate phenotypes has only been seen with one other phase variable protein identified to date, the Class 1 outer membrane protein of *N. meningitidis* encoded by the *porA* gene. *porA* has a homopolymeric tract in the promoter region and altered expression is associated with altered transcription, but it differs from *opc* in that the repeat is a homopolymeric tract of Gs and it is located between putative –10 and –35 promoter elements (van der Ende *et al.*, 1995). In *opc* the homopolymeric tract is located in the expected location of the –35 promoter component and there is no identifiable –35 consensus sequence.

The observation of intermediate phenotypes has been interpreted by Sarkari *et al.*, (1994) to suggest that superimposed upon the transcriptional regulation associated with the number of cytidines, additional (unknown) factors can down regulate Opc expression. Also, it has been suggested that variation in the postulated distinct regulatory site may alter the expression of 'outlier' intermediate phenotypes and affect phase variation rates (Sarkari *et al.*, 1994). However, the mechanism by which alteration in the length of the homopolymeric tract mediates altered transcription is unknown. Possibilities include that the specific sequence of the homopolymeric tract, or a secondary structure that it forms, at certain sequence lengths interact directly with the RNA polymerase complex, either with or without the involvement of additional accessory proteins, thus acting as a –35 element. Alternatively, the repeat may function as a variable spacer within the promoter to alter the relative distance or helical facing of the –10 region and another 5′ region involved in protein binding to the promoter region. The principal purpose of this project was to determine the mechanism by which alteration in the repeat length mediates phase variation of *opc* in a model serogroup B strain of *N. meningitidis*.

7.2 Opc in serogroup B *N. meningitidis* strain MC58

When this study was initiated the only sequenced *opc* gene was from a serogroup A strain C751 (Olyhoek *et al.*, 1991). I wished to study serogroup B *N. meningitidis* because this group of organisms poses the greatest challenge to vaccine development due to the unsuitability of its polysaccharide capsule for vaccine development. The most studied serogroup B strain is MC58 and it is epidemiologically related to strains associated with invasive disease in Europe and elsewhere. So that the results of this work can be easily integrated into the existing body of information on MC58, this strain was selected as a model serogroup B strain in which to perform these studies. To confirm its suitability as a model strain the *opc* gene was cloned, sequenced and compared with the previously published sequence.

Dr. Mike Jennings kindly provided a previously prepared lambda library of *N. meningitidis* strain MC58 (Jennings *et al.*, 1995). A probe for *opc* was prepared from pBE501 (obtained from Dr. Mark Achtman). The *opc* containing insert was excised, used as a template for PCR using primers Opc1 and Opc2 to generate a PCR product representing 557 bp of *opc* coding sequence, labelled and prepared as described in section 2.7.1. This was used to screen the library by Southern hybridisation using a two stage process to identify 14 plaques in the first round and 3 in the second (section 2.9). These were excised from the screening plates and stored in SM buffer.

The plaques from these screens were transformed into *Esch. coli* strain DH5α and 32 ampicillin resistant transformants were selected. Plasmids were prepared and digested with *Eco*RI to excise the inserts, which were of 3 sizes. One of each of these was assessed by restriction digestion with combinations of *Not*I, *Eco*RI, *Cla*I, *Hin*dIII, and *Nae*I which cut within the previously published *opc* sequence and/or the pBlueScript polylinkers. One clone included a small insert of approximately 500 bp, another contained an insert of approximately 3 kb but which did not include the whole *opc* open reading frame and extended from the 3′ end of *opc*. Sequencing the 3′ end of this plasmid revealed a

homologue of an ABC transporter, which lies beyond a homologue of the *dedA* gene of unknown function of *Esch. coli*. The third clone contained an insert of approximately 3 kb containing the whole of *opc* with flanking sequence in both directions. This plasmid, pNJS1, is used in further studies (Figure 2.1).

pNJS1 was sequenced to obtain double stranded sequence of the *opc* gene and flanking sequence using the primers described in section 2.21.1. The sequence is shown in Figure 7.1. This revealed that 5′ of *opc* was a copy of insertion sequence IS1106 and 3′ was a homologue of *dedA*. Between IS1106 and the *opc* promoter region was an additional region of 230 bp that was not present in strain C751. This is shown in figure 7.1 and includes multiple copies of repeat sequences called RS3 (ATTCCC / GGGAAT) and RS2 (which are longer and include RS3 repeats) (Seiler *et al*., 1996). This region was extremely difficult to sequence. Also, the region is sufficiently internally repetitive that primers made within the region led to 'double priming'. This sequence was eventually completed by repeated sequencing of this region with multiple primers, long and short sequencing runs, by increasing the temperature at which the reactions were performed, and a combination of dCTP, deaza-, and dITP sequencing. As indicated by sequence compressions and premature termination sites, the inverted repeat motifs within this region are capable of forming multiple secondary structures. This type of repeated sequence is present in the intergenic regions of the *pilS* locus in *N. gonorrhoeae* (Haas & Meyer, 1986) and is present more abundantly, and in the absence of other repeats thought to be involved in recombination, in the single locus containing the *pilE* and *pilS* genes in *N. meningitidis* strain MC58 (personal observation). The other locations at which this RS2 / RS3 is present suggest that it may be directly involved in recombination processes, but there is no experimental evidence for this at present. These repeats are frequently, but not exclusively, associated with IS1106 insertion sequences (personal observation from the MC58 genome sequence). Their functional significance in this context is unknown.

Figure 7.1. Sequence of *opc* from *N. meningitidis* strain MC58, compared with the previously published sequence from serogroup A strain C751 (Olyhoek *et al.*, 1991) and the sequence generated for serogroup B strain S3446. The major features are labeled. RS3 repeats are underlined. The additional 230 bp sequence which is variably present in different serogroup B strains in the upstream region is in italics. Polymorphisms between MC58 and C751, and between S3446 and MC58 are shown in red. Deletions are represented by full stops. Base 1 is the first base of pNJS1 cloned DNA. The 3′ end of the pNJS1 insert extends to the ABC transporter homologue located 3′ of the *dedA* homologue. The intervening 3′ sequence was not determined. RS3 hexamer repeats are underlined. The termination and initiation codons of the adjacent ORFs are shown in dark green. The IHF consensus sequence is shown in blue and the bases that were altered in pink (section 7.11). The homopolymeric tract, the -10 region, initiation codon and termination codon of *opc* are shown in black bold. The inverted repeat neisserial uptake signal sequence 3′ of *opc* is shown in turquoise.

```
                                                termination codon of IS1106 transposase
          1                                                                          50
opcMC58   gccaacaggt taagtgcgcc cgctgccgcc taaaaggcag cccggatgcc
  C751    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 s3446    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

          51                                                                        100
opcMC58   tgattatcgg gtatccgggg aggattaagg gggtatttgg gtaaaattag
  C751    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 s3446    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

          101                                                                       150
opcMC58   gcggtatttg gggcgaaaac agccgaaaac ctgtgtttgg gtttcggctg
  C751    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 s3446    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

          151                                                                       200
opcMC58   tcgggaggga aaggaatttt gcaaaggtct caaacaaaaa cagaaaccta
  C751    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 s3446    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

          201                                                                       250
opcMC58   aagtcccgtc attcccgcgc aggcgggaat ccagaccccc aacgcggcag
  C751    ~~~~~~~~~~ AATTCCGCGC AGGCGGGAAT CCAGACCCCC AACGCGGCAG
 s3446    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

          251                                                                       300
opcMC58   gaatctatcg gaaataaccg aaaccggacg aacctagatt cccgctttcg
  C751    GAATCTATCG CAAATAACCG AAACCGGACG AACCTAGATT CCCGCTTTCG
 s3446    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
```

```
301                                                        350
opcMC58  cgggaatgac ggcaggtggg tttcagttgc tcccgataaa tgccgccatc
  C751   CGG.AATGAC GGCAGAGTGG TTTCAGTTGC TCCCGATAAA TGCCGCCATC
 s3446   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~ccgataaa tgccgccatc

         351                                               400
opcMC58  tcaagtctcg tcattccctt aaaacagaaa accgaaatca gaaacctaaa
  C751   TCAAGTCTCG TCATTCCCTT AAAACAGAAA ACCGAAATCA GAAACCTAAA
 s3446   tcacgtctcg tcattcccta aaaacagaaa accaaaatca gaaacctaaa

         401                                               450
opcMC58  atccgtcat tcccgcgcag gcgggaatct aggtttgtcg gcacggaaac
  C751   ATTTCGTCAT TCCC~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 s3446   atc..gtcat tcccgcgcag gcgggaatct aggtttgtcg gcacagaaac

         451                                               500
opcMC58  ttatcgggta aaacggtttc tttagatttt acgttctaga ttcccgcctg
  C751   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 s3446   ttgtcgggaa aaacggtttc tttagatttt acgttctgga ttcccgcctg

         501                                               550
opcMC58  cgcgggaatg acgatgaaaa gattgttgtc gcttcggata aatttttgtc
  C751   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 s3446   cgcgggaatg acgatgaaaa gattgttgtc gcttcggata aatttttacc

         551                                               600
opcMC58  gcgttgggtt ctagattccc gcctgcgcgg gaatgacggc ggcgggtttc
  C751   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 s3446   gtgttgggtt ctagattccc ...tgagcgg gaatgacggc ggcgggtttc

         601                                               650
opcMC58  tgttttccg ataaatacac acaaactaaa atttcgtcat tcccataaaa
  C751   ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~ATAAAA
 s3446   tg.ttttccg ataaatacac acaaactaaa atttcgtcat tcccataaaa

         651                    cg g  cg IHF binding site consensus 700
opcMC58  aacagaaaac caagtgagaa taacaattcg ttgtaaacaa ataactattt
  C751   AACAGAAAAC CAAGTGAGAA TAACAATTCG TTGTAAACAA ATAACTATTT
 s3446   aacagaaaac caagtgagaa taacaattcg ttgtaaacaa acaactattt

         701                              homopolymeric tract  750
opcMC58  gttaattttt attaatatat gtaaaatccc cccccccccg aaagcttaag
  C751   GTTAATTTTT ATTAATATAT GTAAAAT..C CCCCCCCCCG AAAGCTTAAG
 s3446   gttaattttt gttaatgtat gcaaaa..at ctcgccccctg aaagttaag

         751 -10                                            800
opcMC58  aatataattg taagcgtaac gattatttac gttatgttac catatccgac
  C751   AATATAATTG TAAGCGTAAC GATTATTTAC GTTATGTTAC CATATCCGAC
 s3446   aatataattg t.agcgtaac gattatttac gttatgttaa tatatccgac

         801                          initiation codon     850
opcMC58  tacaatccaa attttagagg ttttaactat gaaaaaaaca gttttacat
  C751   TACAATCCAA ATTTTGGAGA TTTTAACTAT GAAAAAAACA GTTTTTACAT
 s3446   tacaattcaa attttagagg ttttaactat gaaaaaaaca gttttacat

         851                                               900
opcMC58  gtgccatgat tgccctgacc ggtactgccg ccgctgcaca agagcttcaa
  C751   GTGCCATGAT TGCCCTGACC GGTACTGCCG CCGCTGCACA AGAGCTTCAA
 s3446   atgccatgat cgccctgacc ggtactgccg ccgctgcaca agagcttcaa
```

```
          901                                                       950
opcMC58   accgctaatg agtttaccgt ccacaccgac ctctcttcca tttcttcaac
  C751    ACCGCTAATG AGTTTACCGT CCACACCGAC CTCTCTTCCA TTTCTTCAAC
 s3446    accgctaatg agtttaccgt .cacaccgac ctctcttc.a tt.ct.caac

          951                                                      1000
opcMC58   tcgtgctttc ctgaaagaaa aacacaaagc tgccaaacac atcagcgtgc
  C751    TCGTGCTTTC CTGAAAGAAA AACACAAAGC TGCCAAACAC ATCGGCGTAC
 s3446    tcgtgctt.. ctgaa.gaa. .acac.aagc tgc..aacac atc.gcgtac

          1001                                                     1050
opcMC58   gtgctgatat tcctttgat gccaaccaag gcatccgctt ggaagccggt
  C751    GTGCTGATAT TCCTTTTGAT GCCAACCAAG GCATCCGCTT GGAAGCCGGT
 s3446    gtgctgatat ...ntttgat gccaaccCGgg ggatccgAtt ggaagccggt

          1051                                                     1100
opcMC58   ttcgggcgca gcaaaaaaaa tattattaat ttggaaacag atgagaacaa
  C751    TTCGGGCGCA GCAAAAAAAA TATTATTAAT TTGGAAACAG ATGAGAACAA
 s3446    ttcgggcgca gcaaaaaaaa tatatttaat ttggaaacag atgagaacaa

          1101                                                     1150
opcMC58   gctgggtaag actaaaaatg taaaactgcc caccggcgtt cctgaaaacc
  C751    GCTGGGTAAG ACTAAAAATG TAAAACTGCC CACCGGCGTT CCTGAAAACC
 s3446    gctgggtaag actaaaaatg taaaactgcc caccggcgtt cctgaaaacc

          1151                                                     1200
opcMC58   gtatcgatct ttacacaggc tacacctaca cccaaacgtt aagtgattct
  C751    GTATCGATCT TTACACAGGC TACACCTACA CCCAAACGTT AAGTGATTCT
 s3446    gtatcgatct ttacacaggc tacacctata cccaaacgtt aactcattct

          1201                                                     1250
opcMC58   ttaaatttcc gtgtgggtgc cggcttgggt tttgaatctt caaaagacag
  C751    TTAAATTTCc GTGTGGGTGC CGGCTTGGGT TTTGAATCTT CAAAAGACAG
 s3446    ttaaatttcc gtgtgggtgc cggcttgggt tttgaatctt caaaagacag

          1251                                                     1300
opcMC58   cattaaaacc accaagcata cgcttcacag cagccgtcag tcgtggttag
  C751    CATTAAAACC ACCAAGCATA CGCTTCACAG CAGCCGTCAG TCGTGGTTAG
 s3446    cattaaaacc accaagcata cgcttcacag cagccgtcag tcgtggtCag

          1301                                                     1350
opcMC58   ccaaagttca cgcggatttg ctttcccaac tgggtaacgg ctggtatatc
  C751    CCAAAGTTCA CGCGGATTTG CTTTCCCAAC TGGGTAACGG CTGGTATATC
 s3446    ccaaagttca cgcggatttg ctttcccaac tgggtaacgg ctggtatatc

          1351                                                     1400
opcMC58   aacccttggt ctgaagtgaa atttgacctc aattcccgct ataaattaaa
  C751    AACCCTTGGT CTGAAGTGAA ATTTGACCTC AATTCCCGCT ATAAATTAAA
 s3446    .acccttggt ctgaagtgaa atttgacctc aactcccgc. ataaattaaa

          1401                                                     1450
opcMC58   caccggcgtt accaatctca aaaaagacat caatcaaaaa accaacggct
  C751    CACCGGCGTT ACCAATCTCA AAAAAGACAT CAATCAAAAA ACCAACGGCT
 s3446    caccggcgtt accaGtctca aaaaagacat caatcaaaaa accaacggct

          1451                                                     1500
opcMC58   ggggctttgg attgggtgca aatattggta aaaaactggg cgaatccgcc
  C751    GGGGCTTTGG ATTGGGTGCA AATATTGGTA AAAAACTGGG CGAATCCGCC
 s3446    ggggctttgg attgggtgca aatattggta aaaaactggg cgaatccgcc
```

```
          1501                                                      1550
opcMC58   agcatcgagg cggggccgtt ctacaaacaa cgcacttaca aagaatccgg
  C751    AGCATCGAGG CGGGGCCGTT CTACAAACAA CGCACTTACA AAGAATCCGG
 s3446    agcatcgagg cggggccgtt ctacaaacaa cgcacttaca aagaatccgg

          1551                                                      1600
opcMC58   cgagtttagt gtaacaacca agagtggcga cgtatcgctc accatcccga
  C751    CGAGTTTAGT GTAACAACCA AGAGTGGCGA CGTATCGCTC ACCATCCCGA
 s3446    cgagtttagt gtaacaacca cgagtggcga cgtatcgctc accatcccga

          1601                                  termination codon 1650
opcMC58   aaaccagtat tcgtgaatac ggcttgcgcg tcggcataaa attctgatga
  C751    AAACCAGTAT TCGTGAATAC GGCTTGCGCG TCGGCATAAA ATTCTGATGA
 s3446    aaaccagtat tcgtgaatac ggcttgcgcg tcggcataaa attctgatga

          1651                                                      1700
opcMC58   tttgaaatca tcattgtcac tttaaatgcc aaaccgcaaa ctattttggt
  C751    TTTGAAATCA TCATTGTCAC TTTAAATGCC AAACCGCAAA CTATTTTGGT
 s3446    tttgaaatca tcattgtcgc tttaaatgcc aaaccgcaaa ctattttggt

                                                  inverted pair of neisserial
          1701                                                      1750
opcMC58   ttgcggtttt tacgtgaaat gaatttgaat agccgatgcc gtctgaaact
  C751    TTGCGGTTTT TACGTGAAAT GAATTTGAAT AGCCGATGCC GTCTGAAA.A
 s3446    ttgcggtttt tacgtgaaat gaatttgaat agccgatncc gtctgaaa.a

          uptake signal sequences

          1751                                                      1800
opcMC58   tttcagacgg cattttcca atcggttaaa atacgcggtt tgttttttta
  C751    TTTCAGACGG CATTTTTCCA ATCGGTTAAA ATACGCGGTT TGTTTTTTTT
 s3446    tttcagacgg cattttcc. Atcggttaaa atacgcggtt tgttttttta

          1801      Initiation codon of dedA homologue              1850
opcMC58   ggaacgcgcc gtgcttgcct ccgccatcga cttcatcctc catatcgacc
  C751    AGGAACGGGA ATT~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 s3446    ggaacgcgcc gtgcttgcct ccgccatcga cttcatcctc catatcgacc

          1851                                                      1900
opcMC58   aacacctgct cgcgctgtcg gcgcaatacg gtgtgtggat ttatgcgatt
  C751    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 s3446    aacacctgct cgcgctgtcg gcgcaatacg gtgtgtggat ttatgcgatt

          1901                                              1941
opcMC58   ctgtttttga ttgtttttg cgaaaccggc ctgattgttac
  C751    ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 s3446    ctgtttttga ttgtttttg cgaaa~~~~~ ~~~~~~~~~~
```

Note. The differences at the 5′ and 3′ ends of the C751 sequence probably represent the submission of a small amount of poly-linker sequence rather than being an indication of a different sequence location of *opc* in this strain.

The *opc* open reading frame and promoter region sequence was very similar to the previously published sequence from serogroup A strain C751. The DNA sequence of the open reading frames differed by only 2 nucleotides and the overall sequence identity between the MC58 and the previously published C751 sequence was greater than 99% (comparison shown in Figure 7.1). The sequence of the whole cloned region is also 100% identical to that in the recently completed genome sequence of the same strain (personal observation). Subsequent to the completion of this sequence a paper describing the sequence of *opc* from multiple strains was published (Seiler *et al.*, 1996). This includes the serogroup B strain 44/76, the *opc* sequence of which is identical to that of strain MC58 in pNJS1. This sequencing was performed using primers internal to the previously published *opc* sequence (from 93 to 1240 nucleotides of Genbank sequence M80195). Whilst this did not extend as far 3′ or 5′ as the sequence obtained from pNJS1 they did include the repeat region composed of RS2 and RS3 repeats. This RS2 / RS3 region was present in 13 of 28 serogroup B strains with *opc* studies by Seiler *et al.*, (1996). These regions were polymorphic, consisting of six sequence variants containing variable numbers of RS2 and RS3 repeats. Because the sequence in Seiler's study did not extend to the location of the IS1106 in strain MC58 it is not clear whether the presence of this additional sequence is associated with the presence of the insertion sequence. This study was performed using automated sequencing and it is likely that this method avoided many of the compression problems associated with this region encountered using 'manual' sequencing. Comparison of the MC58 *opc* sequence with the others indicated that the gene in MC58 was entirely typical and therefore that MC58 was an appropriate model organism in which to pursue this study.

The published sequence survey of *opc* in diverse strains revealed some unusual features for an antigenic surface protein. Whilst overall 4.2% of nucleotides were polymorphic, only 2.8% of those within the coding region had polymorphisms, and of the 23 sites in the coding region 9 led to no change in the encoded protein. This is a level of diversity more

typical of housekeeping genes than surface proteins. Furthermore, there was no convincing evidence of recombination within the coding sequence, although this could have been a reflection of the low number of polymorphisms available for the analysis. Whether this reflects the recent acquisition of *opc*, is a reflection of functional constraints, or is a reflection of the effectiveness of variable expression in facilitating the evasion of immune responses is unclear. However, surface proteins, whether phase variable or not, are typically far more diverse.

From limited sequencing of promoter regions (section 7.4) one strain, that was not included in the Seiler study, S3446, appeared to be more polymorphic than any of those that were described. Opc25 and Opc26 were used as primers to PCR amplify the whole of this apparently unusual S3446 *opc* gene, and the product was cloned into pCRII and sequenced (see figure 7.1). This revealed that this is actually a dead version of *opc* with several sites of frame shifts within it. The presence of polymorphisms is therefore likely to reflect degeneration of the sequence after the loss of functional Opc. There are only 4 Cs at the site of the homopolymeric tract, however there is additional sequence at this site which might potentially function as a –35 region. If this were the case then this may have represented sequence from an *opc* gene that was not phase variable in this strain. A non-phase variable *opc* has not been reported to date.

The presence of a 'dead' version of *opc* is conceptually challenging. *opc* is not essential for viability or virulence since it is frequently not expressed and many virulent strains do not have *opc*. However, if *opc* has been recently acquired then it would be expected to confer a significant selective advantage upon the strains which possess it, or it would not be expected to have become as widely prevalent in the population as it is. Spread within the population must either involve clonal expansion and diversification or transfer of *opc* to strains that do not already possess this gene by natural transformation. It might be expected that the repair of a functionally important degenerate gene may also occur through natural transformation and homologous recombination and that this might be expected to occur

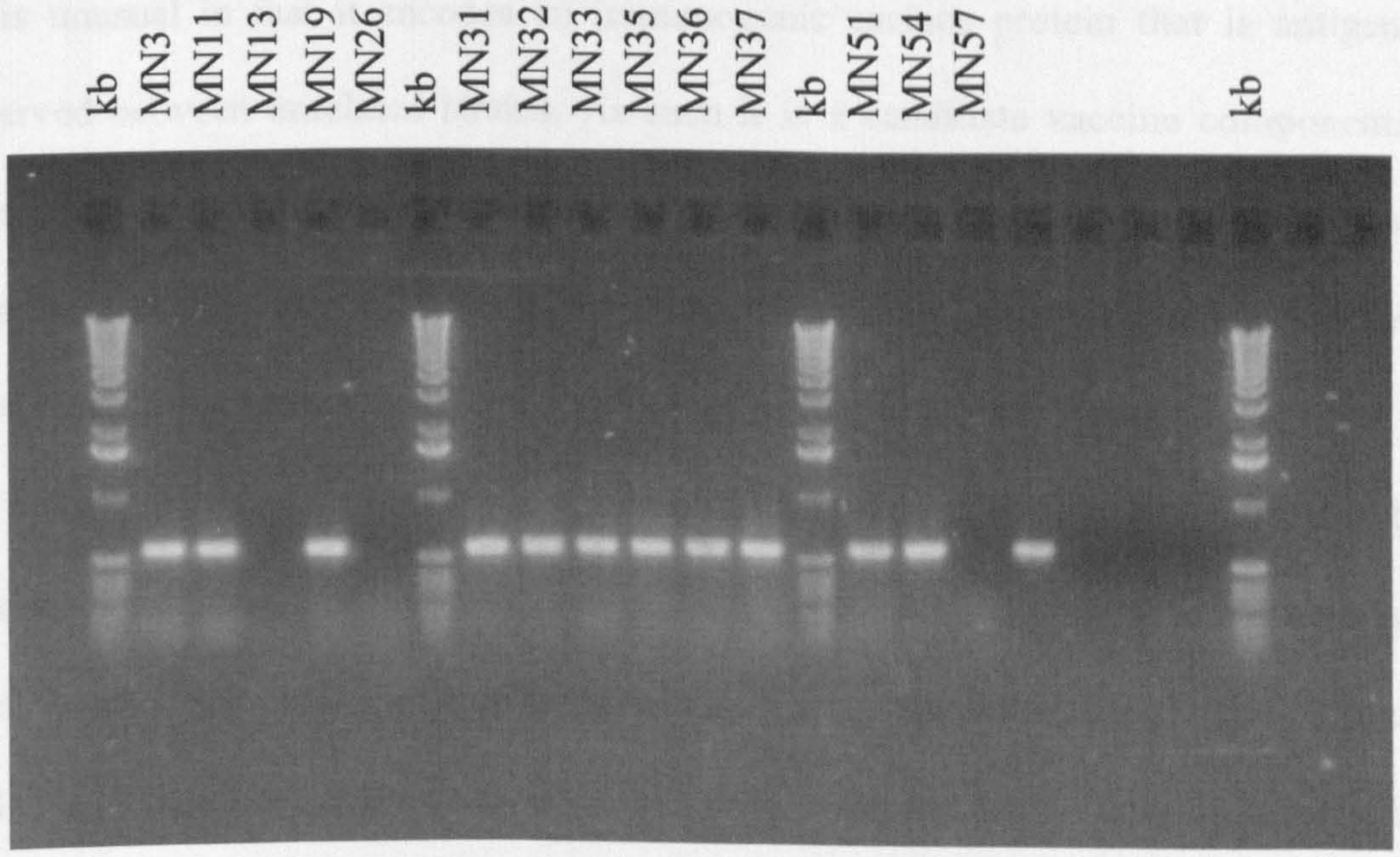
more frequently than the acquisition of a gene for which there is no homologous sequence already present. This has not happened for the *opc* gene in strain S3446. This is the only known example of an *opc* gene with multiple mutations and will make an interesting substrate for experiments to investigate the capacity of horizontal transfer associated with natural transformation to act in gene repair.

7.3 *opc* in serogroup B *N. meningitidis*

When this study was initiated *opc* had almost exclusively been investigated in serogroup A strains of *N. meningitidis*. The presence of *opc* had been reported in 3 serogroup B strains (Sarkari *et al.*, 1994), is present in the large majority of serogroups tested (Achtman *et al.*, 1988; Sarkari *et al.*, 1991) and does not generate the geographical variability seen with other Class 5 proteins (Achtman *et al.*, 1991b). In view of its potential as a candidate vaccine component the presence of *opc* was investigated in a collection of unrelated serogroup B strains. It was decided to use two sources of serogroup B strains: the National Meningococcal Reference Laboratory in Manchester, and the WHO Reference Laboratory in Oslo, Norway. Representative, genetically diverse, epidemiologically unlinked strains were requested from both sources (section 2.1.4).

PCR screening of the Manchester strains using oligonulceotides Opc1 and Opc2 (which amplify a 557 bp section of the *opc* coding region) indicated 11 of 14 positive strains (MN3, MN11, MN19, MN31, MN32, MN33, MN35, MN36, MN37, MN51, MN54) and 3 negative strains (MN13, 26 and 57) (figure 7.2). The same analysis was performed with the epidemiologically unrelated and multi-locus enzymatic electrophoretically different WHO strains, 5 of which (001, BC4, BZ157, M984 and P63) were *opc* negative and 7 (179/82, 44/76, HF130, M470, M986, S3032, and S3446) were *opc* positive by PCR. PCR negative strains and all of the WHO strains were checked with repeat PCR using Opc6 and Opc10, which amplified a 427 bp region of the *opc* coding region with a downstream end 37 bp 3'

Figure 7.2. Ethidium bromide stained 0.8% agarose gel showing PCR products using primers Opc1 and Opc2 to screen strains for the presence of *opc*. This gel shows the results for the Manchester strains with 11 positive (+) and 3 negative (-) results. +C and –C represent positive and negative controls respectively, and kb indicates the size markers. The positive control (+C) is strain MC58, the negative controls (-C) have no primers, no template, and template from strain C111 known not to possess *opc*.



of the Opc1/Opc2 product. The results using both sets of primers were concordant (data not shown).

An *opc* probe was prepared by cloning the PCR product using primers Opc1 and Opc2 into pT7blue. The insert was checked by DNA sequencing, was excised, purified, and labelled with digoxigenin (section 2.7.3). DNA prepared by the CTAB method from each of the Manchester and WHO strains was digested with *Eag*I (which does not cut *opc*) and *Cla*I (which cuts *opc* once) and hybridised in Southern analysis using the *opc* reading frame probe. This is generates a single hybridising band in the *Eag*I digest and two bands in the *Cla*I digest. The results of the Southern blots were concordant with those from the PCR screening, as described above in this section (data not shown).

When this survey was initiated all of the limited number of serogroup B strains in which the presence of *opc* had been investigated in the published literature possessed the gene. *opc* is unusual in that it encodes an immunogenic surface protein that is antigenically conserved between unrelated strains. As such it is a candidate vaccine component. This survey demonstrates that *opc* is not universally present in serogroup B strains, and that whilst it is present in a significant proportion of virulent strains a vaccination strategy directed against it cannot provide protection against all serogroup B strains.

This pattern, in which about 60% of serogroup B strains possess *opc*, was confirmed by the subsequent publication of the survey of *opc* carriage in a large collection of *N. meningitidis* strains in which 28 of 50 unrelated serogroup B strains were found to possess *opc* (Seiler *et al.*, 1996). The higher prevalence amongst the Manchester strains may be a reflection of the sample size, or that there is less genetic diversity and they do not represent a collection reflective of population diversity.

7.4 Investigation of the association between homopolymeric tract length and expression of Opc.

The previous work on *opc* had described a correlation between Opc expression and the number of Cs in the homopolymeric tract. In addition, variation in a single strain between strong and intermediate expression had been associated with changes between 13 and 14 Cs (Sarkari *et al.*, 1994). I wished to confirm the reported correlation between expression and tract length, to determine whether there were other differences in the promoters of unrelated strains that might contribute to the level of expression, and to ensure that strain MC58 was an appropriate model strain in which to investigate the control of *opc* expression.

Each of the *opc* containing WHO strains were investigated for expression of *opc* using serial dilutions of cell extract and anti-Opc monoclonal antibody B306 (kindly provided by Dr. Mark Achtman). In addition, the promoter regions of the WHO strains were sequenced using direct sequencing of biotinylated PCR products using Opc20 and Opc21 primers. This was complicated by the presence of the RS2 / RS3 containing region which led to the generation of larger than expected PCR products that could not be sequenced in some strains. After this additional region had been defined in pNJS1 (as described in section 7.2) it was concluded that the location at which this additional sequence is inserted is likely to be at or beyond the 5′ limit of the functional promoter region. A new 5′ primer (OpcPro) which started immediately 3′ of the RS2 / RS3 region was prepared accordingly and used in combination with Opc16 to amplify 166 bp products of the promoter regions that were cloned and both strands sequenced (section 2.10). The results of these experiments were consistent with the previously described correlation between homopolymeric tract length and expression of Opc. Strains M470, 179/82, and HF130 had 7, 9 and 10 Cs, respectively, and none expressed detectable Opc. Strains MC58 and S3032 expressed high levels of Opc and had 12 and 13 Cs, respectively. Strain 44/76 had 14 Cs and expressed a low level of Opc. Strain S3446 has 4 Cs but this is not associated with an intact *opc* gene so cannot be

interpreted in this context (described in section 7.2). No differences in the promoter regions other than the altered numbers of Cs were identified which might account for the different levels of Opc expression. Notably the sequences of strains 44/76 and MC58 were identical except for the difference of 2 Cs in the homopolymeric tracts. This was considered to be a sufficient basis from which to pursue the mechanism of homopolymeric tract mediated phase variation.

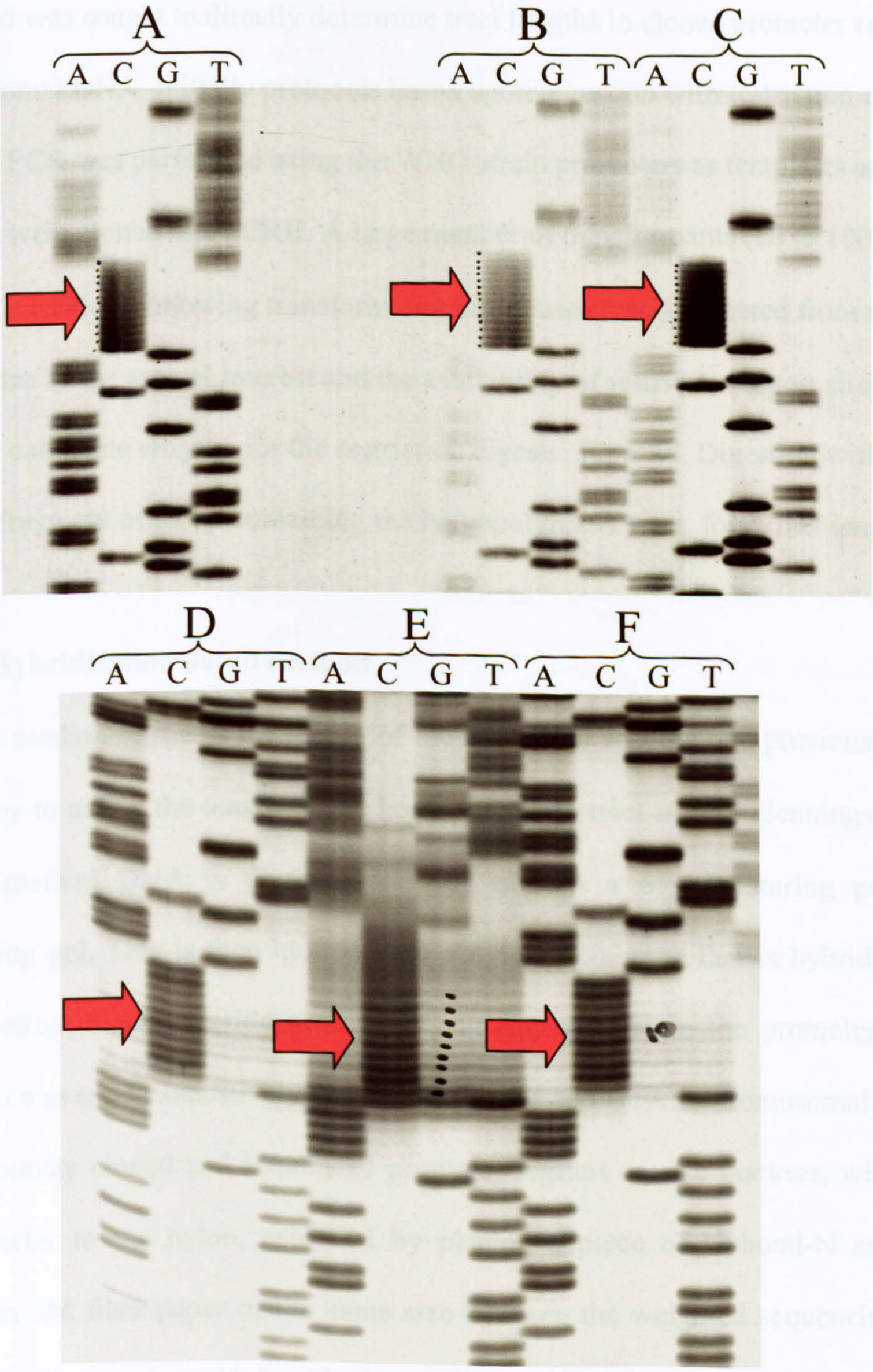
7.5 *In vitro* instability in the homopolymeric tract

During the investigation of homopolymeric tract length described in the previous section significant practical problems, which were related to variability in the length of the homopolymeric tract during experiments, were encountered. When the promoter regions were sequenced directly from PCR products (section 7.4) this yielded results that were frequently difficult to interpret due to variation in the length of the homopolymeric tract in the sequence (figure 7.3). This was a particular problem when the repeat was longer than 10 bases in length and increased as the repeat length increased. This phenomenon can be seen in gels published in other studies of homopolymeric tracts (Sarkari *et al.*, 1994; Jennings *et al.*, 1995). This problem did not occur when sequencing the repeat region directly from the plasmid pNJS1. In addition, during the cloning of the *opc* promoter regions from the WHO strains there were instances when the cloned promoter had a different repeat length from the starting number. One of two cloned promoters from pNJS1 had only 10 Cs instead of 12, one of two cloned promoters from strain M470 had 8 Cs instead of 7. DNA sequence from the pNJS1 template was easy to read either side of the repeat and therefore, as far as it was possible to determine, there was no significant template variability.

Instability in the length of the homopolymeric tract is the source of phase variable expression of Opc. It was not clear to what extent the instability that was observed *in vitro* was due to problems related to direct sequencing, PCR, or to variation in the templates

Figure 7.3

Examples of sequencing gels (read from bottom to top) in which it can be seen that sequence beyond the homopolymeric tracts and the length of the repeat becomes impossible to determine accurately. Sequence templates used in A, B, C and E contained 12, 13, 12 and 14 Cs respectively and the sequences beyond the repeat is indistinct. In contrast, the templates used in D and F contained 10 C s and the sequence remains readable. The red arrows indicate the location of the homopolymeric tracts.



being used, either within chromosomal DNA from neisserial cells or cloned plasmids. This was pursued for two reasons: First, the instability raised important concerns as to the reliability of the tract length determinations. Second, investigation of the variability of these sequences *in vitro* may provide insights into the mechanisms and behaviour of repeat tract variation *in vivo*.

7.5.1 Restriction based approaches

A method was sought to directly determine tract lengths in cloned promoter regions and in chromosomal DNA. Initially protocols based upon digestion with restriction enzymes were pursued. PCR was performed using the WHO strain promoters as templates and the products were cloned into pCRII. A large number of transformants (50 to 100) were selected at random following transformation and plasmid was prepared from each. The limited size of the area of interest and the availability of restriction digest sites presented only one candidate enzyme for the restriction digests: *Tsp*509I. Digestion with this enzyme yields a fragment of 52 bp containing the homopolymeric tract, for a tract length of 12 bp.

7.5.2 Hybridisation based method

The first method to assess the length of the fragments was the one previously used in the laboratory to assess the length of the homopolymeric tract in *lgtA* (Jennings *et al.*, 1995). In this method DNA is digested and run out on a 6% denaturing polyacrylamide sequencing gel. This is then blotted onto a nylon membrane that is hybridised against a γ^{32}P labelled oligonucleotide probe (HPT-Probe) specific to the promoter region. This approach was tried exhaustively with both plasmid and CTAB chromosomal digests, using the previously cloned and sequenced promoter regions as size markers, without success. The transfer to the nylon, achieved by placing a piece of Hybond-N and a piece of Whatman 3M filter paper of the same size between the weighted sequencing plates, was poor. Whilst signal could be obtained when using high concentrations of plasmid a

sufficient signal could not be obtained with the chromosomal preparations. More importantly however, there was never sufficient reproducibility of the position of the internal size standards or the position of DNA fragments from similar promoters running in adjacent lanes for the results to be reliably interpreted. The probable source of these difficulties was the size of the fragment being assessed. The small fragments ran close to the dye and buffer/salt fronts and there was unacceptable consistency in their mobility when compared across the horizontal axis of the gel. This problem was not encountered with *lgtA* probably because the fragment being investigated was 137 bp in length. Despite having persisted with this approach for some time, this method proved not to be sufficiently reproducible. It was however possible to see that there were variations in some of the cloned plasmid promoters, including another example of a change from 12Cs to 10Cs (data not shown). This plasmid was checked by sequencing and had predominantly 10 Cs, with a minor component of 12 Cs also present (no detectable 11 Cs sequence).

7.5.3 Silver staining method

The next approach to this problem used altered gel running conditions and avoided DNA transfer to nylon membranes and oligonucleotide probing. Plasmids containing the cloned promoter regions were digested with *Tsp*509I. These were run out on PAGE gels and silver stained (sections 2.13 and 2.14). Silver staining using the 'Quicksilver' staining kit and protocol was insufficiently sensitive to detect the bands of interest and was associated with excessive background staining when development was prolonged. Alternative silver staining methods were tried. The most sensitive found was 'modification 1' of the method of Johansson and Skoog as described by Vari and Bell (Vari & Bell, 1996). Several of the parameters of this protocol were altered to see whether sensitivity could be increased further. Background staining on development could be reduced and hence sensitivity increased by increasing the washing following the glutaraldehyde sensitisation step from 2, 15 minute washes to at least 6, 10 minute washes. The final protocol is described in section

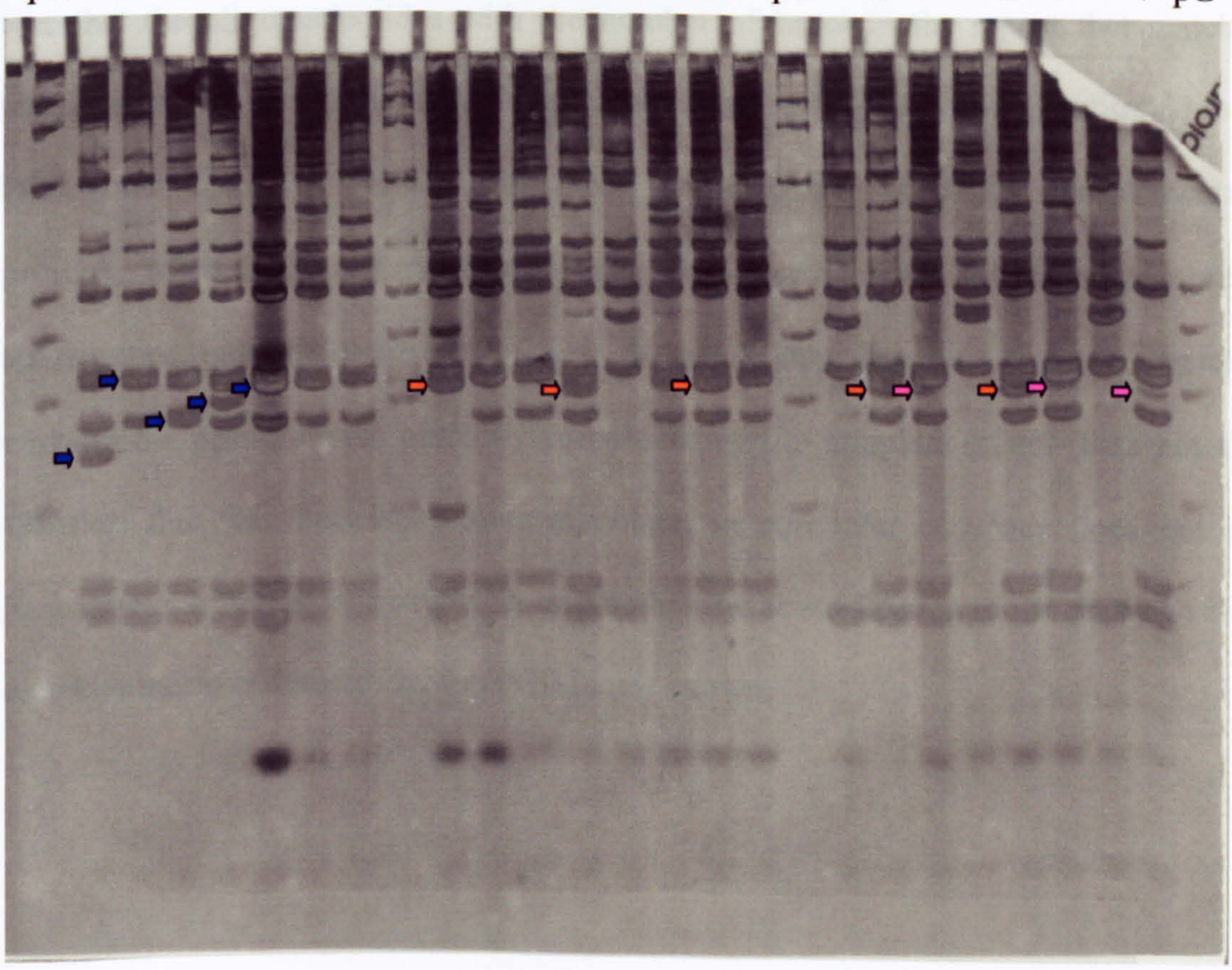
2.14. An example silver stained gel of cloned promoter restriction digests is shown in figure 7.4. The utility of this method was limited, digests frequently contained products of more than one size, and changes of a single base tended to result in broadening of the associated band rather than 2 clearly separable bands. Alterations in the gel composition could not sufficiently increase the resolution of these DNA fragments. However, it was possible to draw some conclusions from the analysis of plasmid digests using this method. These experiments revealed that there were several instances in which the repeat length present in the clone differed from the starting template used in the PCR, and also that some of the transformed cell lines contained plasmids with more than one repeat length present (shown in figure 7.4). It was not possible therefore to determine whether length variation was occurring prior to the cloning and transformation, whether it occurred through instability of the cloned repeat in the transformants, or both. This approach did permit a non-sequenced based analysis of the cloned promoter repeat tract lengths that had been used as templates (and were used as size markers as shown in figure 7.4) in these experiments. The lengths of the repeats as determined by the restriction digests were consistent with those from sequencing (section 7.4). Furthermore, the digests of the plasmids used as starting templates in these and subsequent experiments showed the presence of only a single repeat length in each case.

Problems with the silver staining approach were that whilst sensitivity could be increased sufficiently to assess restriction digests prepared from plasmid preparations of cloned promoters it could not be readily applied to the analysis of chromosomal digests. A biotinylated oligonucleotide was obtained to the highly conserved section of the promoter region (HPT-Probe) previously used in Southern blotting. This oligonucleotide was used to extract the region of the promoter containing the repeat from the *Tsp*509I digests of the promoter region from plasmid and chromosomal DNA (section 2.15b). It was possible to obtain silver stainable bands using the plasmid digests but not the chromosomal preparations. Calculations indicated that the efficiency of extraction from the chromosomal

Figure 7.4 Example of a silver stained gel showing the results for an experiment comparing the lengths of homopolymeric tracts in cloned PCR products using a starting template with 13 Cs.

pG indicate lanes containing pGem molecular weight markers. Lanes labelled 6, 14, 8, 10 and 13 contain digests of cloned promoters with homopolymeric tract of these known lengths which are used as size markers. The fragments containing the repeat are indicated with the small blue arrows. Δ indicates a lane in which there is a major band which differs in length from the template DNA (small orange arrows). + indicates a lane in which there is a minor additional product which differs from the template length (small pink arrows). The degree of variability in the tract lengths in the plasmids indicates either that there was substantial variability in the lengths of the repeats in the cloned material or that there is instability in the repeat after cloning. The presence of multiple fragment lengths from single populations indicates that cells are either transformed by multiple clones which can contain different tract lengths, or that the repeats after transformation are unstable. The latter possibility is less likely, since the plasmids used as size markers consistently have only a single detectable fragment containing the homopolymeric tract.

pG 6 14 8 10 13 pG Δ Δ − Δ pG - Δ + − Δ + − + pG



preparations would have to be high to yield a sufficient quantity of material to visualise by staining with silver (staining sensitivity approximately 1.5ng of DNA). In addition, the separation of the single stranded DNA had to be performed under denaturing conditions and this led to an unacceptable reduction in band resolution on the 15 cm length, 1.5 mm thickness gel format that was used for the previous restriction digest separations. Attempts were made to transfer this methodology to a sequencing gel format using a GelBond PAG (FMC BioProducts, Rockland, ME, USA) gel support and AcrylAide® (FMC BioProducts, Rockland, ME, USA) during silver staining. This resulted in silver staining of the gel-GelBond interface and this approach was abandoned.

7.5.4 Detection by PCR primer annealing

An additional, and ultimately unsuccessful, method that was tried was to use a range of oligonucleotides with different numbers of the repeated bases in PCR reactions. The theory was that the oligonucleotides with the number of bases in the repeat that matched the template would have higher annealing temperatures and that this could be titrated to give a PCR product only when the primer and template were perfectly complementary. In practice, when PCR was performed using 1°C steps around the calculated annealing temperature, products were obtained at the same annealing temperatures regardless of whether the primers were perfectly complementary to template or not (data not shown). Whether this was because primer mismatch in particular sequence does not impede annealing as much as was expected, or whether it was a reflection of the presence of mixed template that was detected by the PCR is not known.

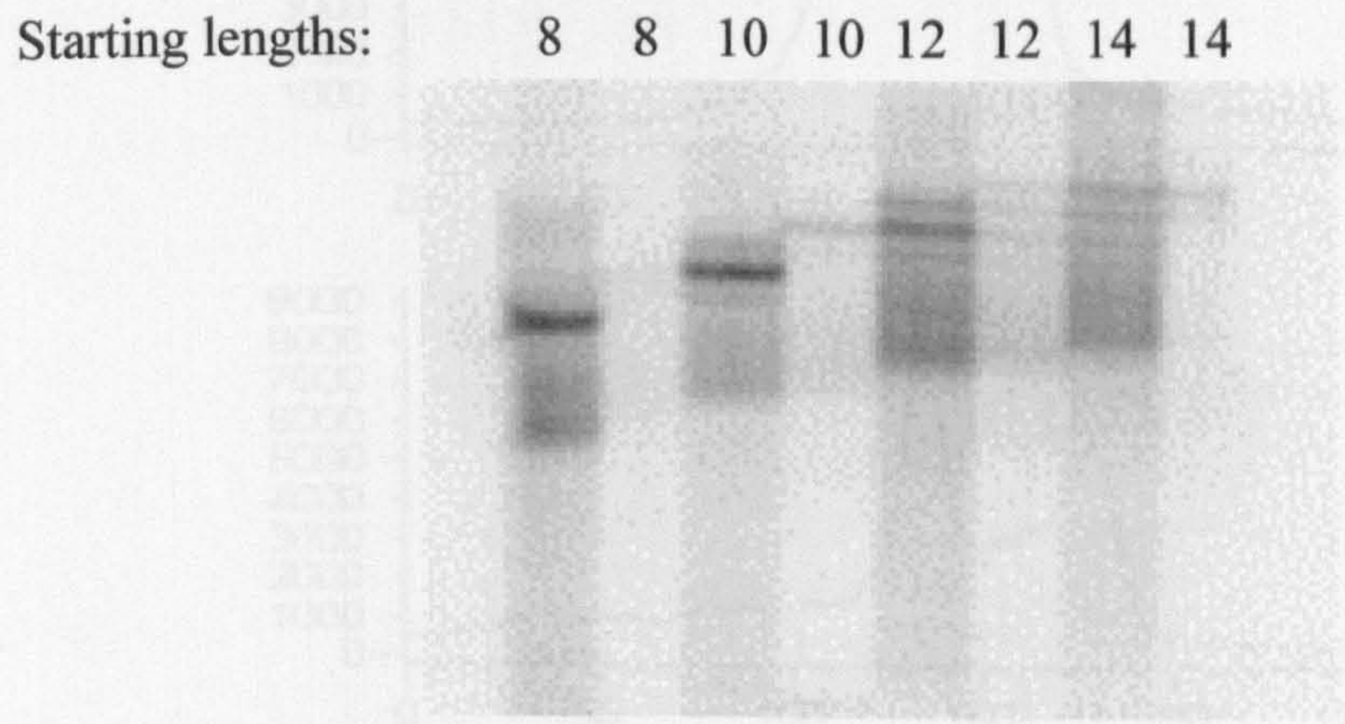
7.5.5 Detection of length variation by direct labelling of PCR products

I was unable to design a method for the direct determination of the homopolymeric tract length from the chromosome. However, a methodology for the direct determination of the variation during PCR amplification was developed (section 2.15). In summary this used PCR amplification of the promoter region and digestion of the PCR product with an enzyme (*Hin*dIII) that had a single T as the first base of the overhang. The digested PCR product was labelled with a reaction mixture that contained ^{35}S dATP as the only nucleotide. Because of the single T base in the overhang any labelled strand could only have incorporated a single radiolabelled base. The labelled strand that was analysed was the strand that started with the primer, so there would be no length variation due to the presence of variable length poly-A tails at the 3′ end. In this way all detectable variation in the length of the resolved labelled product could be assumed to reflect alterations in the length of the hypermutable homopolymeric tract. Examples of the results of these experiments are shown in figure 7.5. When gels were imaged using the phosphorimager it was possible to quantify the tract length variation that occurs *in vitro* (figure 7.6).

The experiments were performed using starting templates containing 8, 10, 12 and 14 Cs and revealed that there was substantial variation in the length of the PCR amplified products that was sufficient to account for the observed variance in tract lengths during direct sequencing of PCR products and in promoters cloned using PCR amplification steps (figure 7.5). The size range and the amount of variant length product tended to increase with the length of the repeat but was still visible in the products obtained using the shortest template of 8 Cs. In some instances the predominant product had a repeat length different from that of the starting template (figure 7.6). Notably, the products did not appear to occur with a normal distribution around the starting template and in some instances there was a predominance of products that had changed by two bases rather than one (figure 7.6).

Figure 7.5 Examples of gels showing length variation in the homopolymeric tract during PCR amplification.

Homopolymeric tract length variation was determined by digestion and end labelling of the PCR products (using OpcPro and Opc16). In the upper gel, pairs of products from separate PCRs using the same template are shown (second set less intense than the first). The two reactions using a starting of 10 Cs in the upper gel clearly show that the predominant products sometimes differ when using the same starting template. The lower gel shows a similar gel adjacent to the sequence of the promoter (using Opc16) used to determine tract lengths . Variability can be seen to increase with the length of the homopolymeric tract in the template promoter, this is particularly clear in the upper gel.



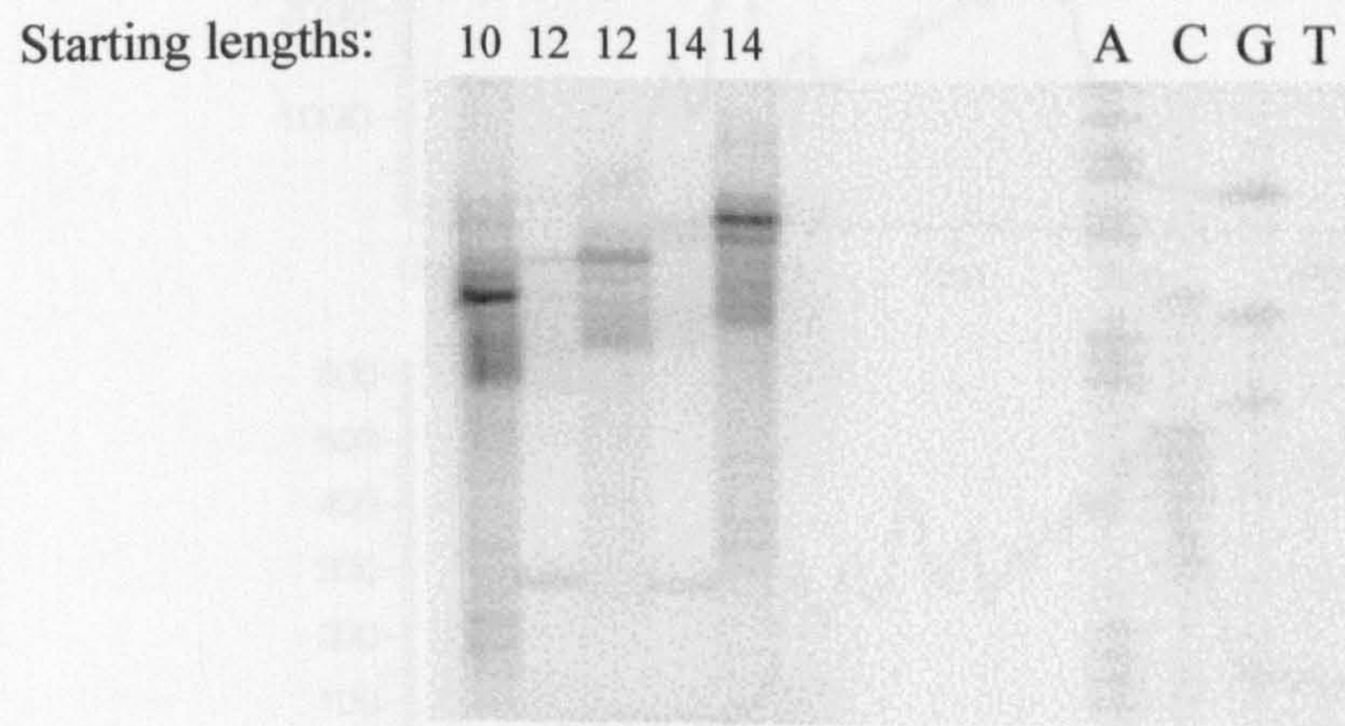
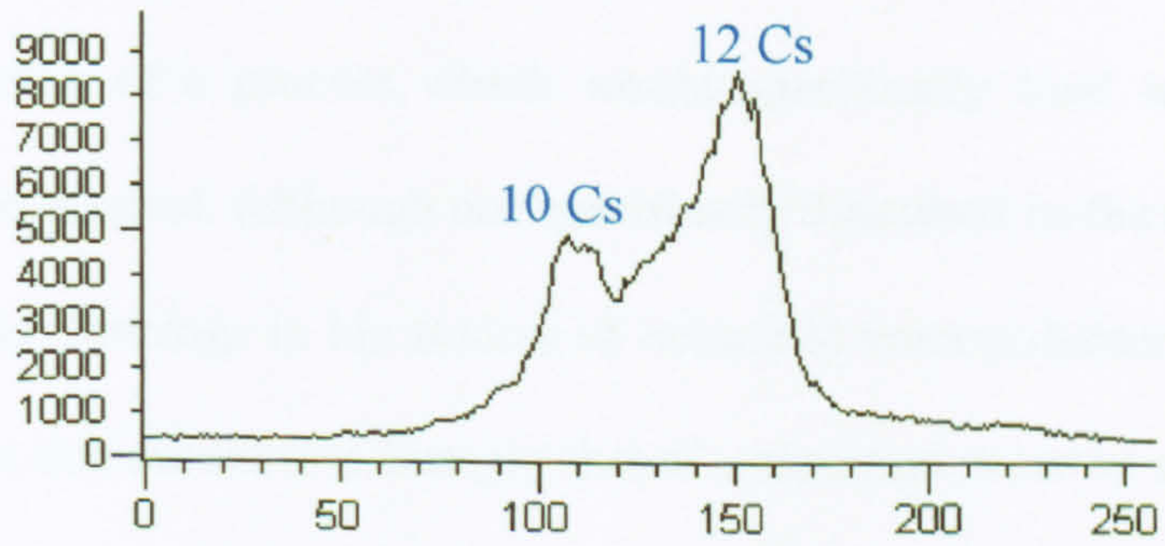


Figure 7.6

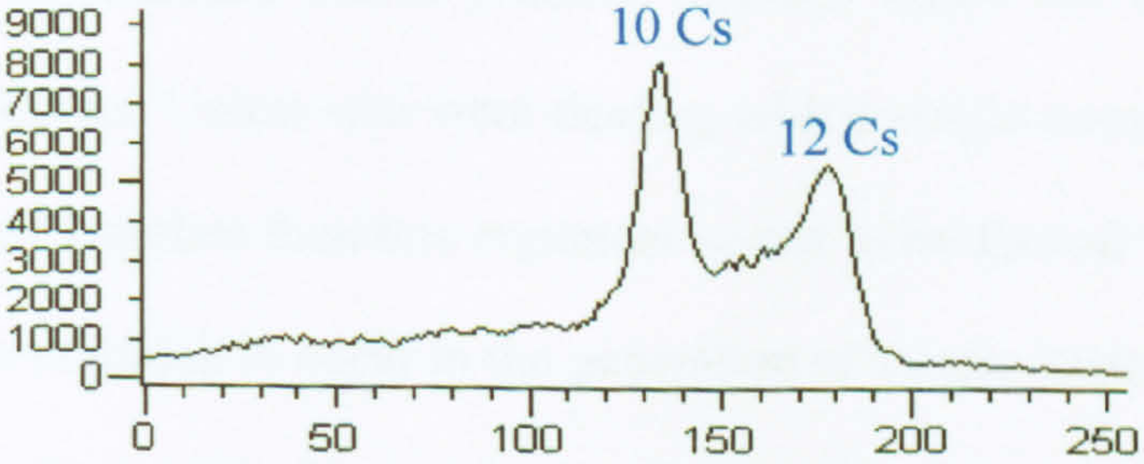
Examples of phosporimager traces analysed in ImageQuant quantitating the changes in homopolymeric tract lengths occurring during PCR.

A and B show different reactions using the same starting template (10 Cs). In both cases there is an abundant peak that is 2 bases higher than in the template sequence. In A the variant sequence is actually more abundant than the product representing the starting sequence. In C and D examples using starting templates with repeats of 12 and 14 Cs respectively are shown. In these cases substantial variation generating a non-normal distribution of product lengths can be seen. The scales are arbitrary and assigned by the ImageQuant software.

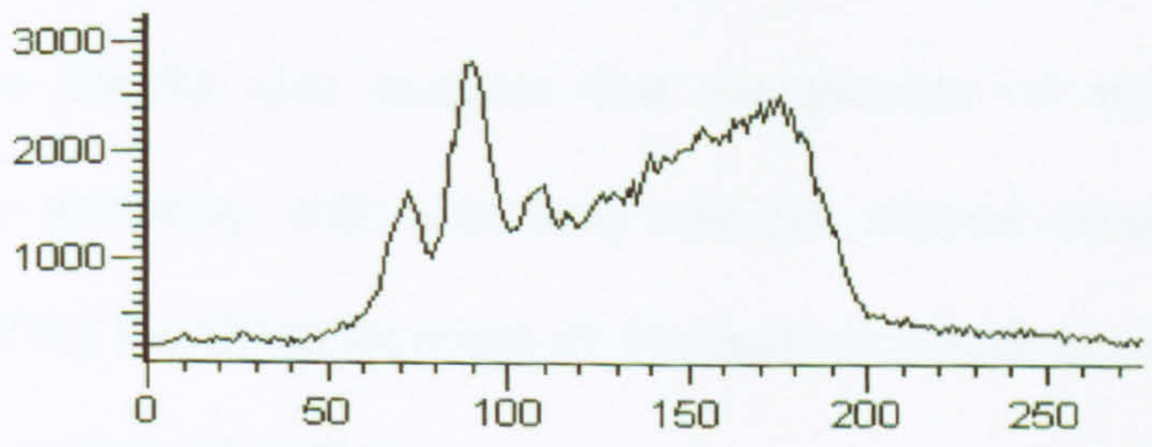
A.



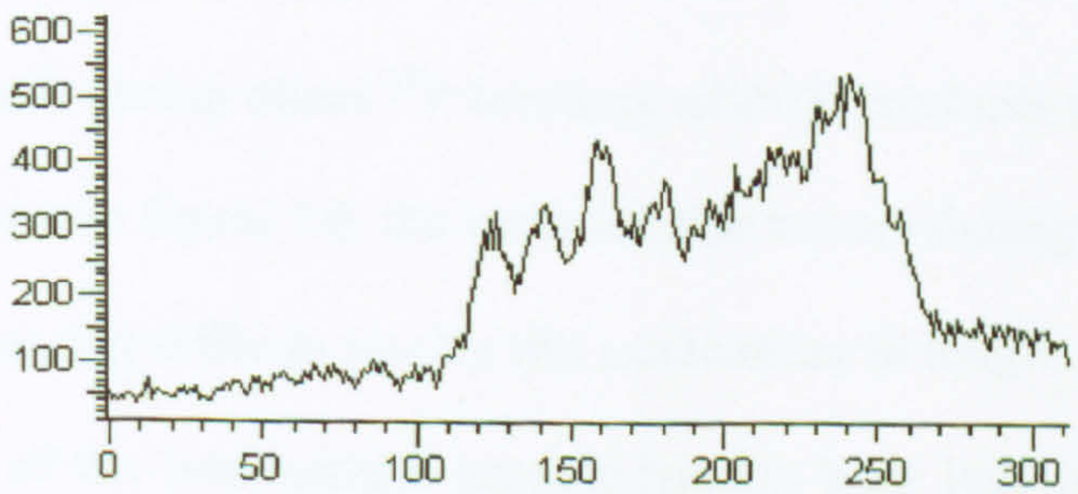
B.



C.



D.



7.5.6 Discussion of *in vitro* repeat instability results

One of the motives for the analysis of *in vitro* variation of the tract was the difficulty associated with sequencing that has involved PCR. PCR dependent variation seems to occur at a sufficiently high rate to account for the problems encountered when sequencing and also the occurrence of repeat lengths in cloned PCR products that differ from those in the starting template. It has been argued that the length of this type of homopolymeric tract can be reliably determined by dividing PCR reactions and then pooling the product prior to sequencing, or sequencing the products separately. This approach assumes that the variation that occurs during PCR amplification occurs due to early stochastic 'founder effects' and in the absence of a process which would specifically tend to increase or decrease the length of the product. Although not specifically described in the methods, this method was used by Mike Jennings in his studies of neisserial homopolymeric tracts (e.g. Jennings *et al.*, 1995). In this model it is thought that if a mutation were to occur early in one reaction leading to an abundant altered product, that this would not be expected to occur in the parallel reactions. Unless one were dealing with a single copy (or at least a very small number) of the template then this argument seems to be flawed since it would be just as unlikely for the mutation to occur in the generation of the products from multiple templates in a single reaction as it is from multiple reactions. The use of parallel PCRs in determining repeat tract lengths also assumes that the process of variation during amplification has some similarity with that proposed for slipped strand mispairing (Levinson & Gutman, 1987), involving increases or decreases in length at similar rates, by a single repeat unit (i.e. one base at a time).

The results of the experiments using direct ^{32}P labelling of PCR products suggest that this model is incorrect. As shown in figure 7.6, the variation that occurs during PCR can result in an abundance of products that differ in size by two nucleotides in length. Whether this is due to greater instability of the intermediate homopolymeric tract lengths or whether it reflects a slippage process that preferentially alters tract lengths by multiples of 2 bases is

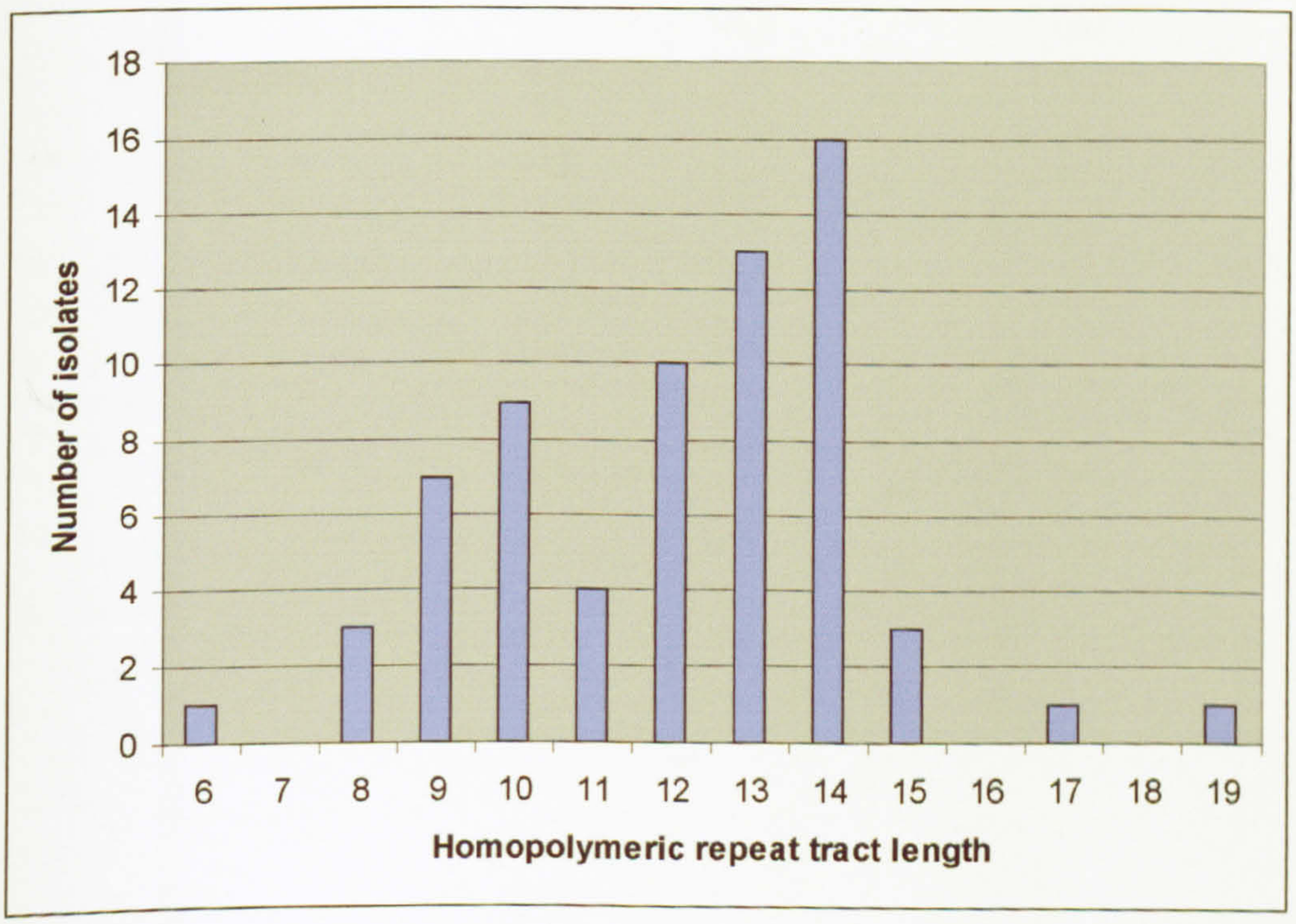
not known. In addition, the process appears to be directional, because the altered products do not occur as a normal distribution around the starting template number.

Finally, and most striking, there are instances in which the predominant product differs from that of the plasmid template (template length determined by sequencing and silver staining of restriction digests of plasmids). Theoretically this should never happen (Bull & Pease, 1995). Mutations do occur during PCR and can result in the cloning and sequencing of a PCR generated mutation and PCR can be used to deliberately introduce mutations. However, when many copies of the starting template are present the predominant base at any position in the product should always reflect the template. This is because a mutation that occurs in one round of amplification from one template is unlikely to occur at the same site in all the parallel replications of the other templates. The pattern of *in vitro* variation demonstrated by the direct labelling of PCR products suggests that there is a novel, apparently directional, high rate process of polymerase slippage that occurs during the amplification of this homopolymeric tract.

If one considers that *opc* should be intermittently expressed at high levels, perhaps during early colonisation of the nasopharynx, and that variation in the homopolymeric tract is not directional, or in other words is likely to increase or decrease in length at similar rates, then an even distribution either side of the numbers associated with strong expression (12 and 13) would be expected. If the rate of variation increases with the length of the homopolymeric tract then this might result in the distribution being slightly skewed towards shorter repeat lengths. This is because if longer repeats vary more quickly they will be more unstable, and hence more likely to reduce in length again than a shorter repeat is to increase in length. Figure 7.7 shows the combined distribution of tract lengths described in this thesis and the larger set of promoter lengths from different strains that have been published previously (Sarkari *et al.*, 1994). This reveals that the lengths of the reported homopolymeric tracts is not the even distribution either side of the numbers

Figure 7.7

Graph showing the distribution of homopolymeric tract lengths from *opc* promoters from isolates of *N. meningitidis*. Note the non-normal distribution and the relative lack of repeats of greater than 14 bases in length. As reported by Sarkari *et al.*, (1994), the 15 and 13 columns include one isolate each in which the phenotype was discordant with the number of bases they determined in the homopolymeric tract.



associated with Opc expression (12 and 13 Cs) that would be expected on the basis of the assumptions described above.

In figure 7.7, the most abundant tract length is 14 Cs and there are very few repeats of 15 Cs or greater. Looking at just the numbers associated with intermediate expression the bias to a distribution there are more isolates with 14 C s than there are with 11 Cs (16 with 14 Cs and 4 with 11 Cs). The observed distribution is one in which there is a progressive decrease in homopolymeric tract lengths from 14 to 8 Cs in length, with a tendency for an abundance of repeat lengths with even numbers of bases. There is a critical repeat length associated with the ability of this type of repeat to form triplex structures that lies between 14 and 18 bases (Kohwi-Shigematsu & Kohwi, 1991). Formation and consequent instability of such unusual DNA structures may account for the lack of repeats of greater than 14 bases in length. This may combine with a process of tract length variation similar to that revealed by the direct labelling of PCR products to account for the observed distribution of homopolymeric tract lengths. This presents a possible alternative model for tract length variation that differs from that of simple gain or loss of the repeated base leading to fluctuation in repeat tract numbers about those associated with expression. Instead, poly-C and poly-G repeats, above a threshold length, may increase in length due to polymerase slippage, only to reduce again when a critical length of greater than 14 bp is reached due to instability.

The instability of the homopolymeric tract during amplification and cloning may explain why the previously cloned *opc* gene from the group A strain C751 (Olyhoek *et al.*, 1991) contained 10Cs when obtained from an Opc positive phenotype, the C751 promoter region being otherwise identical to that of strain MC58 in which strong expression is associated with 12 Cs (figure 7.1). It does not address the issue of why a repeat length of 10 Cs, which would not normally be associated with expression in *Neisseria*, leads to expression in *Esch. coli*. Although this observation highlights the importance of studying bacterial promoters in the correct metabolic and chromosomal environments.

Slipped strand mispairing is routinely invoked as the mechanism that underlies repeat instability, including homopolymeric tracts, in phase variation (Moxon *et al.*, 1994; Jennings *et al.*, 1995; Hammerschmidt *et al.*, 1996b; Chen *et al.*, 1998; Lewis *et al.*, 1999). As the name describes, 'slipped strand mispairing' is a process that involves transient relative strand slippage and subsequent mispairing. This can occur during local denaturation of the DNA strands either during DNA replication or during transcription, or simply as 'strand breathing', which can occur during stationary phase. These misaligned sequences generate structures that lead to the addition or deletion of repeated units by processes that normally mediate DNA repair (Streisinger & Owen, 1985; Levinson & Gutman, 1987). This is a process that is likely to be predominant in mediating changes in repeats other than homopolymeric tracts (Murphy *et al.*, 1989; Belland *et al.*, 1989) but it is clearly distinct from polymerase slippage that can also occur during replication. The observed variability in the length of the repeat in the *opc* promoter during PCR suggest that polymerase slippage rather than slipped strand mispairing may be the source of homopolymeric tract instability. That polymerase slippage may be a mechanism, or perhaps even the predominant mechanism, mediating this variability has important implications with respect to the factors that might influence the rate and nature of phase variation *in vivo*.

7.6 Site directed mutagenesis of the promoter to investigate the mechanism by which the homopolymeric tract length affects expression

The homopolymeric tract is unstable, making the use of templates that contain the native repeats difficult or impossible to use in studies of gene expression and DNA – protein interaction. To overcome this problem a series of site directed mutations was generated in the promoter region. In these mutants the homopolymeric tract was replaced by a number of differing lengths of spacer sequence (see figure 2.1). These included a range of sequence lengths covering those associated with full expression (12 and 13 bases), intermediate

phenotypes (11 and 14 bases) and 'off' phenotypes (10 and 15 bases). In addition, some constructs equivalent to greater than the naturally occurring repeat sequence lengths were prepared (up to an equivalent of a homopolymeric tract length of 24 bp). The purpose of these extended sequences was to insert an extra helical turn into the sequence because some upstream promoter elements can demonstrate return of function when helical facing is preserved even when spacing is increased (Gaston *et al.*, 1990). The sequence that replaced the homopolymeric tract was composed entirely of G and C bases so that the local melting temperature of the promoter was not altered. Once the number of bases in the replacement tract was equal to those of homopolymeric tracts associated with expression (i.e. greater than 13, except for the replacement tract equivalent to 14 Cs which was entirely composed of Gs and Cs) then the spacer sequence was no longer restricted to G and C bases.

This was done by amplifying the region of pNJS1 from the *Hin*dIII site, immediately 3′ of the homopolymeric tract, to the polylinker 5′ of *opc* (using the primers described in section 2.21.4 and LT7) and cloning the product into pNJS1 (Figure 2.1). Because pNJS1 contains two *Hin*dIII sites, one in the vector polylinker and the other in the *opc* promoter, the vector was linearised by performing a partial digest with *Hind*III, the linearised plasmid was purified and then digested with *Eco*RV, and the plasmid used for ligation was obtained by purification of the larger fragment following the second digestion. Initially, the PCR product was cut and ligated into the location of the equivalent section in pNJS1, but very few transformants, and in many cases none, were obtained using this approach. Subsequently the PCR products containing the altered repeat tracts were cloned into pCRII. The altered region was checked by sequencing and restriction mapping to ensure that the expected changes were present at the previous site of the homopolymeric tract (data not shown), and these cloned altered promoter regions were used for the construction of the altered pNJS1 plasmids. The altered promoter regions were cloned into pNJS1 using the *Kpn*I and *Hin*dIII restriction sites from the pBluescript polylinker (in pNJS1 and in the

cloned insert) and in the *opc* promoter respectively. Using this approach plasmids containing replacement tracts equivalent to homopolymeric tracts of 10, 11, 12 and 13 Cs were obtained. These were checked by PCR amplifying the promoter region and cloning into pCRII (using Opc16 and OpcPro) and sequencing the whole promoter region (data not shown). However, using this approach plasmids containing homopolymeric tract replacements equivalent to lengths greater than 13 Cs could not be obtained and all final clones, except with the 13 C replacement tract, had restriction digest patterns or sequencing information indicating that other changes had occurred during construction. The reason for the inefficiency of this approach is unknown.

A new approach that avoided the use of partial digests was pursued. The pBluescript polylinker *Hin*dIII site in pNJS1 was removed by digestion with *Eco*RV and *Hin*cII and the product re-ligated. The cloned upstream regions containing the altered promoters were then cloned into the equivalent location of pNJS1 using *Kpn*I and *Hin*dIII. Once the hompolymeric tract region had been replaced, a kanamycin cassette from puc4kan was inserted immediately 3′ to the inverted repeat neisserial uptake signal sequence downstream of *opc* (figure 7.1) in each plasmid. The cassette was located in this position to leave the *opc* gene intact and to avoid locating the cassette 5′ of *opc* where it would have the potential to influence promoter function. This was achieved by partial digestion of the pNJS1 plasmids containing the sequences replacing the homopolymeric tracts with *Bsa*AI to linearise them (there is a second site in the vector sequence), dephosphatasing the linearised plasmid, and ligating a *Hin*cII blunt ended kanamycin cassette at this site. Transformants were selected on LB plates containing 50 mg/l of kanamycin, plasmid preparations were made, and the site of kanamycin cassette insertion was determined by restriction digestion with *Xba*I. Clones with the kanamycin cassette in the correct position were selected and named pC10R, pC11R, pC12R, pC13R, pC14R, pC+3R, pC+5R, pC+7R, pC+9R and pC+11R, that contained replacement tracts equivalent to 10, 11, 12, 13, 14, 16, 18, 20, 22 and 24 bp respectively.

Each of the kanamycin resistant plasmids containing the altered promoter regions were used to perform transformations of the acapsulate *N. meningitidis* strain MC58 ∝3 (Virji *et al.*, 1995). Kanamycin resistant transformants were selected on Leventhals plates containing 50mg/L of kanamycin. Kanamycin resistant colonies were subcultured and screened by PCR, using a primer specific for the sequence replacing the homopolymeric tract (RepScreen) and OpcPro to amplify the promoter region. As the length of the replacement promoter sequence increased above 13 there was decreasing efficiency of successful replacement of the homopolymeric tract (the recipient strain had 12 Cs in the sequence being replaced). Eventually replacements using all plasmids except pC+9R and pC+11R were obtained, but replacements with these longer sequences, equivalent to repeat sequences of 22 and 24 bp, could not be obtained despite screening several hundred kanamycin resistant transformants. It is assumed that the transformation was occurring leading to the gain of kanamycin resistance but with a cross-over forming between the kanamycin cassette and the *opc* promoter so that the promoter region was not exchanged. As the difference in length between the homopolymeric tract being replaced and the replacement tract increased, the less frequently was this section included in the recombination event.

Kanamycin resistant transformants of strain MC58 that were positive in PCR screening were subcultured and colony immunoblots were performed using the anti-Opc monoclonal antibody B306. Initially there were some results that did not accord with the expected correlation between tract length and Opc expression. Some of the transformants using pC12R were Opc negative, some of the transformants with pC+3R and pC+7R expressed Opc and one transformant with pC+7R had a mixed phenotype. In order to identify the cause of this variability and the reason for these unexpected results the promoter regions were amplified using OpcPro and Opc16 and PCR products were sequenced. This showed that the sequenced Opc negative pC12R had deleted a repeated GC dinucleotide from the replacement tract leaving the equivalent of only 10 Cs. The sequenced Opc positive

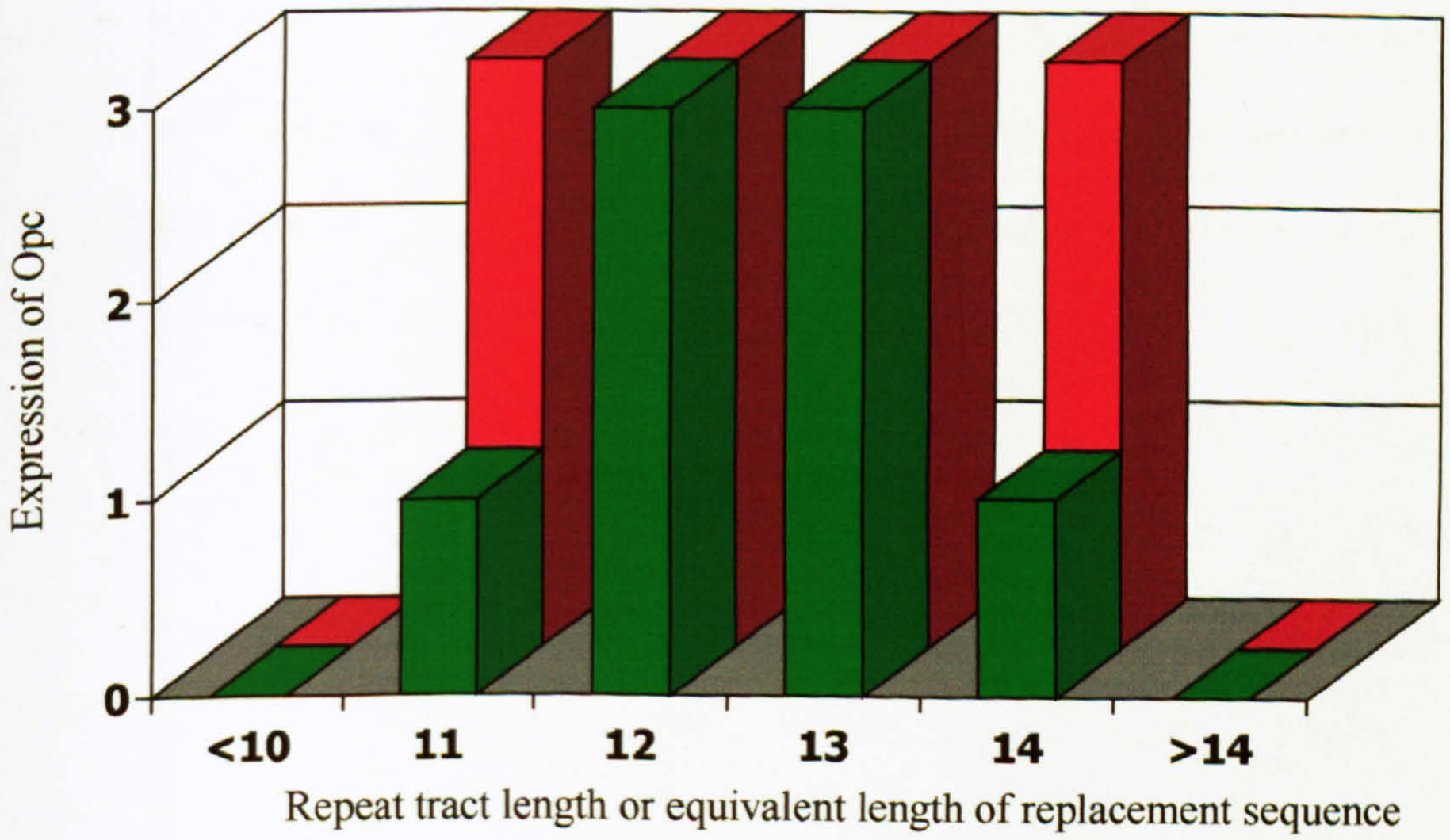
pC+3R transformants had lost the same GC dinucleotide and the pC+7R transformant had in addition deleted a repeated GCCGC pentanucleotide, leaving them both with tracts equivalent to 14 Cs. Since the replacement tracts had been designed so that the local melting temperature would not be changed this meant that a stretch composed entirely of G and C bases had to be used, and this inevitably contained a degree of local repetitivity. Ultimately, transformants were obtained and their phenotypes determined with all of the engineered changes to the repeat region except for with C14R. However, it was possible to use the pC+3R transformant that had deleted a GC dinucletide in its place, since it had an equivalent tract length.

The results of expression of Opc in the cells in which the homopolymeric tract were replaced are represented in figure 7.8. Expression of Opc is associated with replacement tract lengths that are equivalent to those seen in strains with the native homopolymeric tract. However, the intermediate phenotypes no longer occur. Replacement tract lengths equivalent to homopolymeric tract lengths associated with intermediate levels of Opc expression (11 and 14 bp) mediate strong expression of Opc. This indicates that the homopolymeric tract itself is not required for the expression of Opc. In addition, it suggests that there are two processes by which variable repeat length controls expression. The first is a spacing effect, whereby tract lengths of 11 to 14 bp are associated with expression of Opc. The second is an independent process mediated by the specific sequence of the homopolymeric tract that results in the in intermediate phenotypes which is not recapitulated by the replacement tract sequences.

RNA polymerase typically contacts regions of the DNA promoter, which it specifically recognises, termed the –10 and –35 regions (von Hippel *et al.*, 1984; McClure, 1985; Busby & Ebright, 1994). The homopolymeric tract in the *opc* promoter is located in the expected location of the –35 region of the promoter and there is no discernible –35 consensus sequence. It is unlikely that the specific sequence of the homopolymeric tract, or a secondary structure it can form at certain lengths, acts as a –35 in the context of Opc

Figure 7.8

Graph showing the relationship between homopolymeric tract (green) and replacement tract length (red) and the expression of Opc. Opc expression was determined by colony immunoblotting of *N. meningitidis* strains using monoclonal antibody B306. This used strains with a range of natural homopolymeric tract lengths and strain MC58 ∝3 transformed with pC10R, pC11R, pC12R, pC13R, pC14R, pC+3R, pC+5R and pC+7R. Strong expression of Opc leads to rapid development of intense immunostaining, whereas weak expression results in slow development which never becomes intense. This is expressed semiquantitatively on this graph on which 0 indicates no detectable expression, 1 indicates weak expression, 3 indicates very strong expression of Opc.



expression in the site-directed mutants in which the repeat has been replaced. It is well established that the length of the spacer DNA between –10 and –35 regions affects promoter function (Stephano & Gralla, 1982; Mulligan *et al.*, 1985; Ayers *et al.*, 1989). If the variable repeat in *opc* controlled expression through effect on promoter component spacing then this would have some similarity to the variable dinucleotide repeat located between the –10 and –35 regions of the phase variable *hif* genes of *H. influenzae* (van Ham *et al.*, 1993). However, *opc* differs in that the repeat sequence must be influencing the relative position of the –10 and another promoter component that presumably lies 3′ of the repeat and therefore further apart than promoter components currently known to be affected by sequence repeats. The potential for the specific sequence of the bases between –10 and –35 elements in a promoter, as opposed to its length, to affect promoter function has also been investigated (Auble *et al.*, 1986). In these experiments sequences capable of adopting non-B DNA structures were inserted into a bacteriophage promoter and linked to a *lacZ* reporter, including a poly-C homopolymeric tract. These experiments demonstrated that spacer DNA between the –10 and –35 regions in *Esch. coli* can affect the rate of open complex formation *in vitro* and the levels of gene expression from plasmids *in vivo*. Of the sequences studied, homopolymeric tracts of C and G had the greatest effect on expression, reducing expression when an otherwise optimal number of bases were present between promoter components. Experiments with different spacer sequences of similar length indicate that the binding of RNA polymerase is sensitive to the relative orientation of the –10 and –35 regions and that the two regions are contacted simultaneously. Further study of the promoters that contain homopolymeric tracts of C or G supported a model in which these sequences form a stable structure with reduced twist and/or rise when compared with B-DNA, and that the affect on expression in this context is due to an alteration in the rotational orientation (helical facing) of the flanking –10 and –35 regions with respect to each other (Auble & deHaseth, 1988; Warne & deHaseth, 1993). These studies which inserted different sequences between the –10 and –35 promoter components were designed

to show whether it was possible that the spacer sequence could affect promoter function. In pursuing this, the experimenters deliberately made use of artificial promoters that would contain sequence capable of forming unusual structures. The effects of these promoters were then studied in *in vitro* binding studies and in plasmids. In contrast the current study demonstrates, for the first time, a difference between the behaviour of a promoter containing a naturally occurring homopolymeric tract of Cs and that seen when this repeat is replaced with sequence of equivalent sequence length. This is therefore the first example of a facing effect mediated by a repeat sequence influencing promoter function in a naturally occurring promoter and that this occurs in a chromosomal context. In addition, it shows that these effects can extend over a greater sequence distance than previously recognised and can affect the formation of a transcriptional complex involving a –10 region and a sequence at least two helical turns upstream. Furthermore, the separate contributions of spacing and facing are demonstrated by the ON-OFF switching and the presence of the intermediate phenotypes respectively. This is the first instance in which gene expression has been demonstrated to be controlled by a combination of spacing and helical facing mediated by a naturally occurring sequence element.

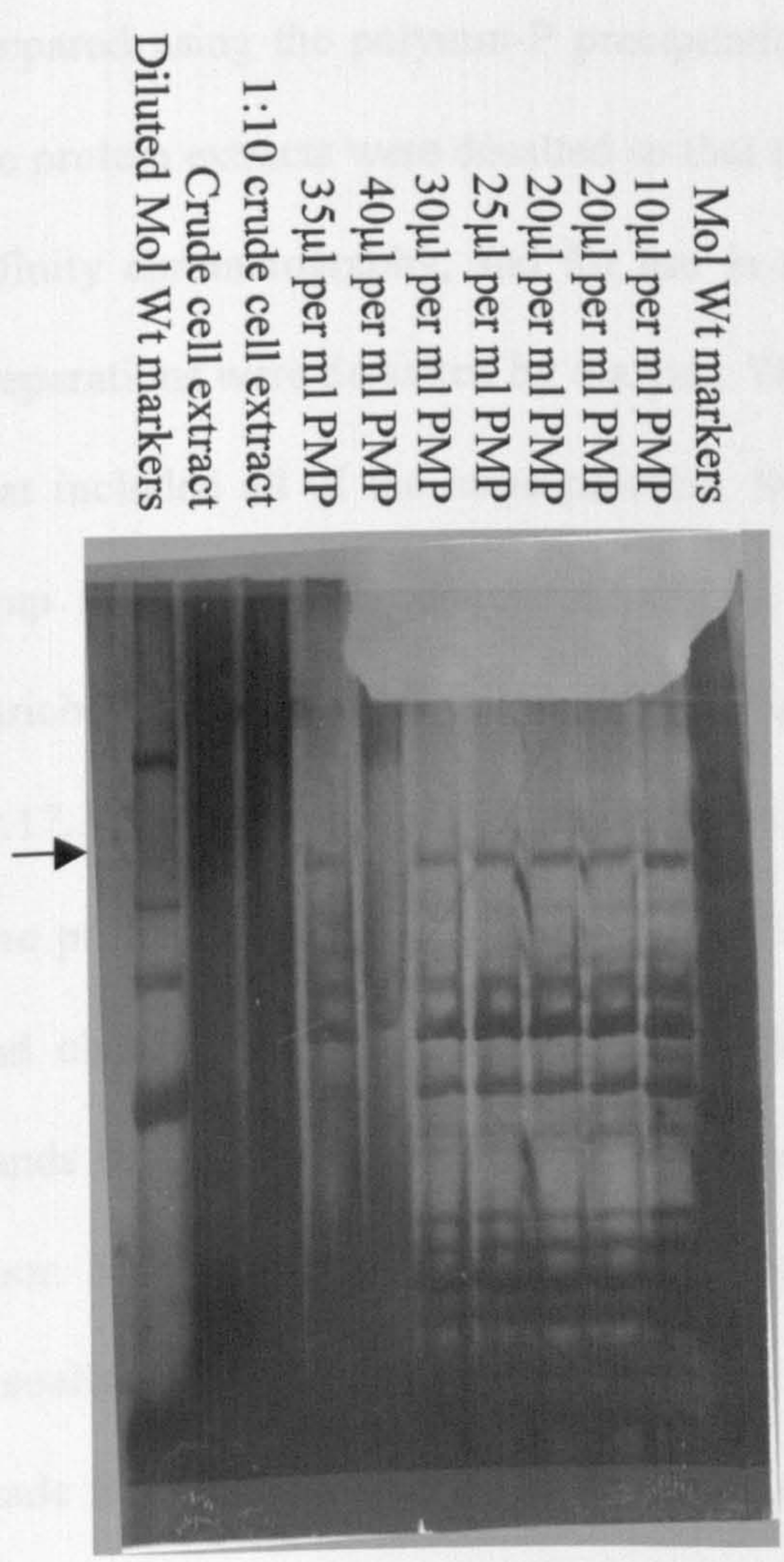
7.7 The purification of *N. meningitidis* RNA polymerase using polymin-P precipitation

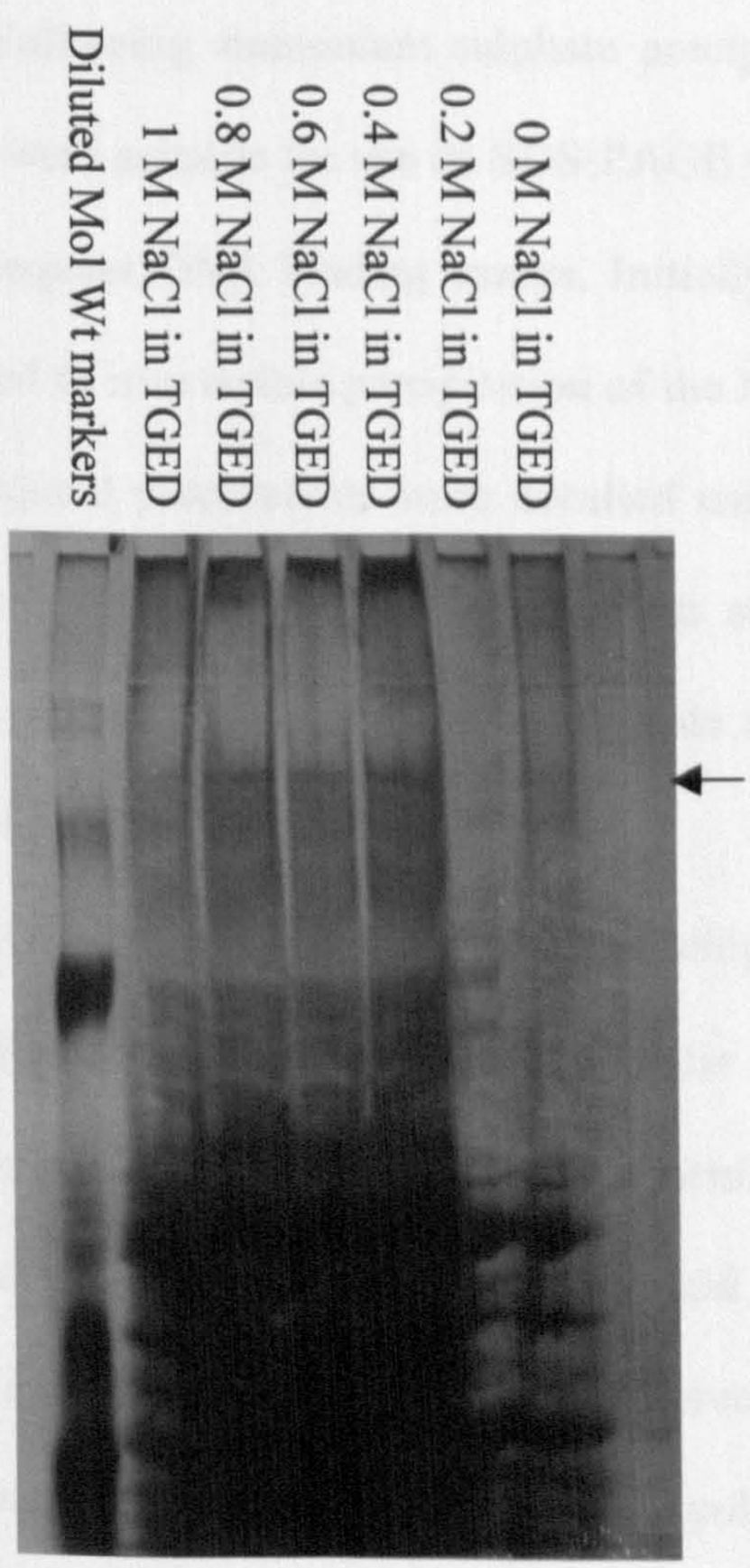
The repertoire of sigma factors in *Neisseria gonorrhoeae* differs from that in *Esch. coli* (Klimpel *et al.*, 1989) and it cannot be assumed that other DNA binding proteins are similar between these *Neisseria* spp. and *Esch. coli*. The number of Cs in the *opc* containing plasmid pBE501 (section 7.1) associated with expression is not consistent with those seen in *N. meningitidis* (Olyhoek *et al.*, 1991) and there are no identifiable compensatory sequence differences to account for this difference. If effects of DNA topology which involve the repeat region are important in the function of the promoter it cannot be assumed that the promoter will act similarly in plasmid and chromosomal

contexts. For these reasons it was decided that wherever possible the promoter function must be investigated using *Neisseria* derived DNA binding proteins and expression should be studied in a *N. meningitidis* chromosomal context.

Attempts were made to purify RNA polymerase from *N. meningitidis*. These studies were performed using strain MC58 ∝11 which lacks capsule, Opc, Opa proteins and pili (Virji *et al.*, 1995) rendering it safer to grow in large quantities for these purposes than working with the wild-type strain. Purification was initially attempted from cell lysates using a primary precipitation using polyethylenamine (polymin-P) (section 2.17.1), which has been used previously to purify RNA polymerase in several species (Burgess, 1969; Burgess & Jendrisak, 1975; Jendrisak & Burgess, 1975; Lowe *et al.*, 1979; Chamberlin *et al.*, 1983; Schneider *et al.*, 1987). Polymin-P is a positively charged polymer in neutral solutions and causes a precipitation of acidic proteins at low ionic strength by forming charge neutralisation complexes and cross bridges between the complexes. When used at appropriate pH and ionic strength it preferentially precipitates DNA binding proteins which can then be eluted with buffer containing increasing amounts of salt (NaCl). Titration of the polymin-P precipitation of cell lysates prepared in TGED was performed with a range of concentrations and the changes in the cell lysate were determined by SDS-PAGE. The loss of high molecular weight bands (presumed to be the ß-subunits of RNA polymerase) and a band of the expected molecular weight of the sigma component of RNA polymerase from the cell lysates was achieved with the addition of 35 µl of polymin-P per ml of lysate (figure 7.9 - left hand gel). This was the same concentration of polymin-P reported to precipitate RNA polymerase from *Esch. coli* cell lysates and was co-incident with the clearing of the lysates, consistent with what was described in the original papers describing the method (Burgess, 1969; Burgess & Jendrisak, 1975). Elution of the precipitated protein using a range of NaCl concentrations in TGED was performed to optimise washing and elution conditions to elute the bound RNA polymerase, and the eluted proteins were examined by SDS-PAGE. Little high molecular weight protein was

Figure 7.9 Silver stained (following Coomassie) SDS-PAGE gels showing the precipitation of proteins from cell lysates from *N. meningitidis* with polymin-P and elution of proteins from the precipitate with increasing concentrations of NaCl. The high molecular weight bands believed to be indicative of RNA polymerase β-subunits are indicated by the arrows.
The left hand gel shows the precipitation of proteins. The lowest concentration at which the high molecular weight bands are partially precipitated corresponds to 30 µl of polymin-P per ml of lysate and there is complete precipitation in with 35 µl of polymin-P per ml of lysate.
The right hand gel shows the elution of proteins from the polymin-P precipitates. The high molecular weight proteins can first be seen with 0.4 M NaCl in TGED.
These gels had to be stained with silver to visualise the high molecular weight proteins which is the reason for the high background and highlights the difficulties encountered when working with these preparations.





eluted at concentrations up to 0.4 M NaCl and greater amounts of the bands thought to represent RNA polymerase subunits were eluted using between 0.6 and 1.0 M NaCl in TGED to wash the precipitates (figure 7.9 - right hand gel). The primary enrichment and purification conditions selected were to wash the precipitate with 0.5 M and elute the putative RNA polymerase with 1.0 M NaCl in TGED. This yielded protein preparation significantly enriched for the large molecular weight proteins of similar molecular weight to RNA polymerase ß subunits.

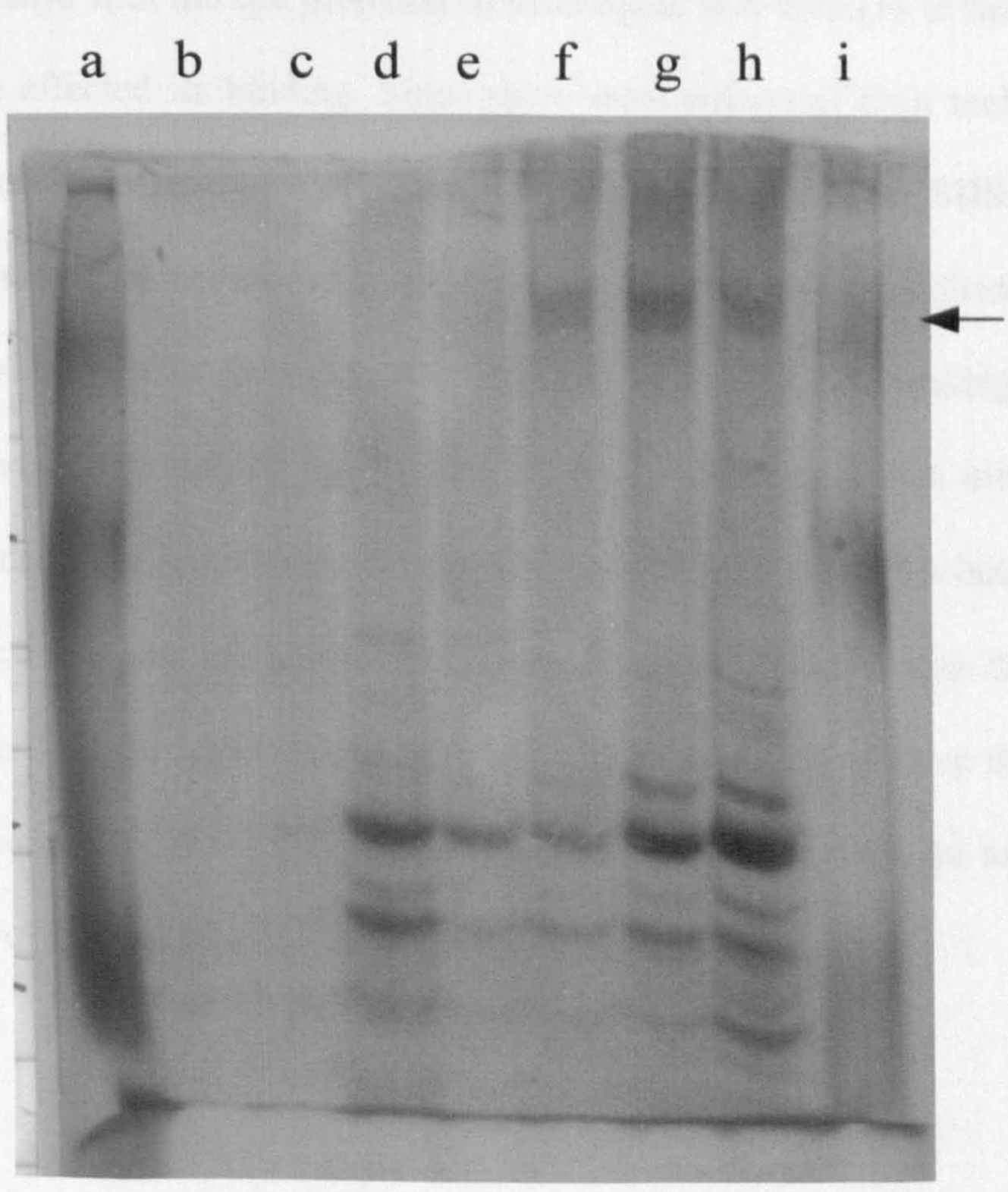
Ammonium sulphate precipitations (2.17.2) were titrated in order to determine the appropriate concentrations needed to enrich for, and concentrate, the high molecular weight bands, using both the whole cell lysates (figure 7.10) and the protein extract prepared using the polymin-P precipitation. Following ammonium sulphate precipitation the protein extracts were desalted so that they were suitable for use on SDS-PAGE gels, in affinity chromatography, and for use in subsequent DNA binding assays. Initially these preparations were de-salted by dialysis. This led to irreversible precipitation of the fraction that included all of the large proteins. Subsequent preparations were desalted using Hi-Trap mini desalting columns (2.17.2b). The polymin-P enriched, ammonium sulphate enriched, desalted material was then affinity purified on a 5ml heparin sulphate column (2.17.3) (Sternbach *et al.*, 1975; Davidson *et al.*, 1979; Lerbs *et al.*, 1985).

The protein extracts, as assessed by SDS-PAGE, following the polymin-P precitipitation and elution steps were not consistent. Specifically the yield of high molecular weight bands thought to represent the large subunits of RNA polymerase was frequently very poor. Silver staining was usually required to see the larger bands which could not be visualised using Coomassie blue staining. In light of this, ammonium sulphate precipitates made from the crude cell extracts were compared with those prepared using the polymin-P precipitation method. SDS-PAGE gel analysis of the resulting extracts showed that the ammonium sulphate precipitation gave a greater yield of high molecular weight proteins than the polymin-P precipitation. The polymin-P prepared protein extract was then

Figure 7.10 Coomassie stained SDS-PAGE gel showing serial ammonium sulphate precipitations of proteins from centrifuged crude cell lysates of *N. meningitidis*.

Lanes show HiTrap column desalted protein precipitates obtained following step-wise increases in ammonium sulphate concentration. Lane a contains kaleidoscope markers. Lane b contains the precipitate obtained with 10% ammonium sulphate, lane c contains the precipitate from the supernatant of the 10% precipitation obtained with 20% ammonium sulphate and so on in 10% increments progressively up to lane g with 60%, and finally lane h shows the precipitate with a 20% increment to 80% ammonium sulphate. Lane i contains diluted kaleidoscope markers.

The putative RNA polymerase β subunits of RNA polymerase of approximately 150kDa are indicated with an arrow.



analysed by Heparin affinity chromatography and in electrophoretic mobility shift assays (EMSA) (section 2.18) to determine whether purified RNA polymerase could be obtained from the extract and whether it contained the expected DNA binding characteristics. The high retention material obtained following heparin affinity chromatography, in which DNA binding proteins are expected to be enriched, did not contain high molecular weight bands typical of RNA polymerase ß-subunits on SDS-PAGE. The EMSA was used to detect the presence of proteins capable of binding the *opc* promoter, using crude cell extract, the polymin-P prepared extract and the high retention material from heparin affinity chromatography. This showed that only the crude cell extract contained proteins which formed low mobility complexes as were expected for a complex formed by RNA polymerase (data not shown). The polymin-P based method was therefore abandoned.

Originally the intent had been to purify the neisserial RNA polymerase and to use it in DNA binding assays with the *opc* promoter to investigate how changes in the repeat region of the promoter affected its binding. Since there were neisserial high molecular weight proteins that could apparently be mistaken for RNA polymerase on SDS-PAGE gels I decided that this was an unsatisfactory method for following the purification process. I therefore decided to monitor the subsequent attempts to purify DNA binding proteins with EMSA using the *opc* promoter, which detects the presence of DNA binding proteins directly by retardation of radiolabelled target DNA mobility in gels. This has two principal advantages over using gels to visualise the purified proteins. First, it uses the presence of the biological activity for which the protein is being prepared to monitor the purification and it is therefore specific for the proteins of interest. Second, it avoids the assumption that the only protein of interest binding to the promoter is RNA polymerase.

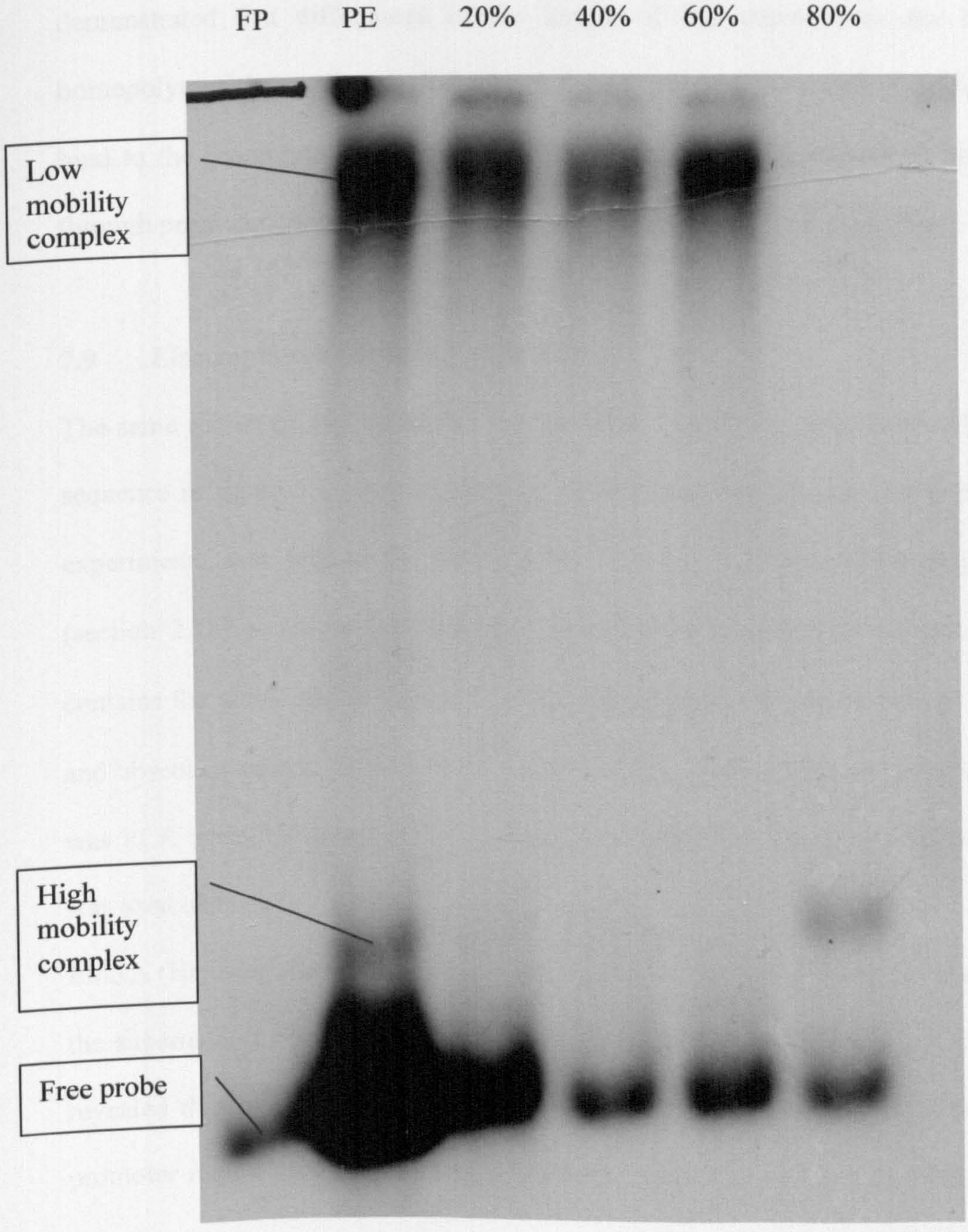
7.8 DNA binding protein extraction using polyethylene glycol precipitation and interactions of protein extracts with the *opc* promoter

An alternative primary enrichment protocol has been described for RNA polymerase which does not use polymin-P. Instead it uses precipitation with, and subsequent elution of DNA binding proteins from, polyethylene glycol (PEG) precipitated DNA (Gross *et al.*, 1976). Use of this method is not widely reported and it seems to have been replaced by polymin-P and specialised column based methods. However, it had some potential advantages. First, it avoids the step of DNA shearing, which is an important step in polymin-P precipitation method and is associated with perceptible heating of the lysate even though it is performed on ice. Secondly, the PEG precipitation method is not dependent upon the affinity of proteins for an added substrate nor is it as sensitive to lysate salt and protein concentrations or pH since it involves precipitation of the DNA to which the proteins are already bound rather than to a second substrate. Thirdly, this was perceived to be an approach that could be extended to the study of other DNA binding proteins. This approach was successful and protein extracts prepared in this way generated similar DNA protein complexes on EMSA to those seen with whole cell extracts (data not shown) (see lane labelled PE in figure 7.11 for appearance with whole cell extract). I therefore decided to use the DNA promoter sequence itself to affinity purify the *opc* binding proteins from the DNA binding protein extract prepared by PEG precipitation and to use this to look at the promoter-bound proteins directly by SDS-PAGE.

Promoter regions to be used as templates were amplified by PCR, cloned and checked by sequencing from pNJS1 and from site directed mutants with replacement tracts equivalent to homopolymeric tract lengths associated with OFF (pC10R), intermediate (pC11R) and ON (pC12R) phenotypes. The promoters were then amplified using a biotinylated Opc16, bound to streptavidin coated Dynabeads, and used as described in section 2.17.5. The binding reaction used the same conditions for binding that had been found to be optimal for protein binding to DNA in the EMSA. This procedure demonstrated the presence of

Figure 7.11.

EMSA using the *opc* promoter region as a probe and with whole cell protein extract from which the RNA polymerase and other proteins has been progressively precipitated using increasing concentrations of ammonium sulphate (i.e. using the supernatant). FP indicates free probe control, PE indicates un-precipitated protein extract, the percentage ammonium sulphate used in precipitations is indicated above the relevant lanes. A smaller complex can be seen in addition to the larger, low mobility, complex in both the crude extract and in the 80% precipitated protein preparation. The faint band at the top of the lanes is material retained in the wells of the gel.



DNA bound proteins including large molecular weight bands consistent with the ß-subunits of RNA polymerase, visible in the affinity purified extract (figure 7.12). These bands were not as abundant in the PEG prepared extract, demonstrating that they had been specifically enriched by extraction on the *opc* promoter sequence. These bands were visible in the extracts prepared with the promoter lengths associated with strong expression (pNJS1 promoter with 12 Cs and pC12R promoter with an equivalent length replacement tract). These high molecular weight bands were not visible with other replacement tract lengths (pC10R and pC11R promoters with replacement sequences equivalent to 10 and 11 Cs which would have OFF and intermediate phenotypes, respectively) (figure 7.12). This demonstrated that differences in the length of the sequence in the location of the homopolymeric tract directly influences the capacity of the putative RNA polymerase to bind to the promoter. This is consistent with the model for control of binding mediated through promoter component facing and spacing described in section 7.6.

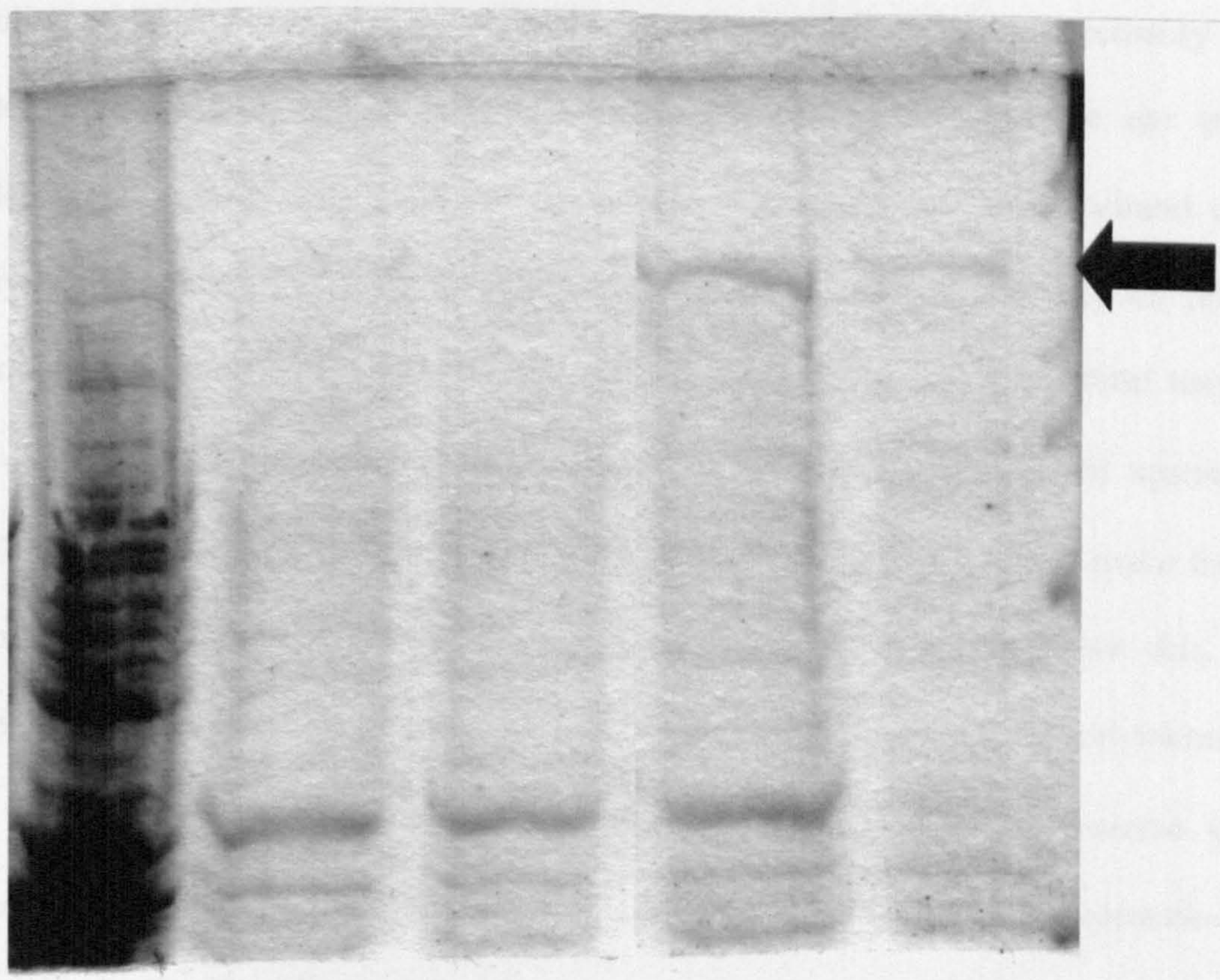
7.9 Electrophoretic mobility shift assays

The same region of 166 bp of the *opc* promoter, which ran from bases 638 to 803 in the sequence in figure 7.1 (as amplified by OpcPro and Opc16), that had been used in other experiments, was used as the template for EMSA (otherwise known as gel shift assays) (section 2.18) to assess the binding of proteins to the promoter of *opc*. This sequence contains the whole of the region 3′ of the RS2/RS3 repeat, that is present in some strains and absent in others. Originally, promoter regions with a homopolymeric tract of 12 Cs was PCR amplified from pNJS1. When it was available the equivalent region of pC12R was used instead.

EMSA (Henninghausen & Lubon, 1987) was then performed using crude cell extracts and the supernatants obtained during ammonium sulphate precipitation (section 2.17.2). This revealed that there were two specific DNA-protein complexes formed between the *opc* promoter region used as a probe and proteins present in cell lysates from *N. meningitidis*

Figure 7.12

SDS-PAGE gel of protein extracts prepared by the PEG precipitation method and then affinity purified on *opc* promoter sequences (sections 2.18.6 and 7.8). The large bands have a molecular weight of approximately 150 kDa (size markers not shown on this gel) and are consistent with RNA polymerase β subunits (marked with the arrow). PEG indicates a lane containing a sample of the protein extract prior to affinity purification, in this lane the promoter-binding enriched complex is not visible as a major component. The complex is clearly visible only in those affinity purifications performed using promoters with tract lengths associated with expression (12 Cs). C10R, C11R and C12R are promoters with stable replacement tracts equivalent to 10, 11 and 12 Cs, 12Cs is a promoter with a native C12 homopolymeric tract.



Note – the gel image has been edited to remove an unloaded lane.

(figure 7.11). One had very low mobility consistent with a DNA – RNA polymerase complex and the other had a much higher mobility suggesting an interaction with a small DNA binding protein (figure 7.11). The presence of the low mobility complex in supernatants from precipitations performed with up to 60% and absent in 80% is consistent with the late precipitation of the putative β subunits of RNA polymerase seen previously (figure 7.10). The presence of the high mobility complex suggests that there is a small DNA binding protein that has a high affinity for the *opc* promoter region.

7.10 Binding of the α subunit of RNA polymerase to the *opc* promoter

Binding of the putative RNA polymerase to the promoter in EMSA and in affinity purification was always done using protein extracts likely to contain other DNA binding proteins. The absence of a candidate –35 promoter element and the presence of the higher mobility DNA-protein complex detected in the EMSA presented the possibility that the binding of RNA polymerase may depend upon the binding of a small accessory protein. Alternatively the protein in the high mobility complex may bind the *opc* promoter independently, potentially acting as a transcriptional regulator. The α subunit of RNA polymerase is known to mediate interactions between RNA polymerase and the promoters through interaction with sites upstream of the –35 region. This interaction may either involve independent binding of the α-subunit of RNA polymerase to the upstream site within the promoter, or it may be mediated by accessory proteins which make the initial contact with the DNA (Ebright & Busby, 1995). In order to investigate this, EMSA experiments were performed using the *N. meningitidis* α-subunit of RNA polymerase

In collaboration with Dr. John Davies the α subunit of RNA polymerase from *N. meningitidis* strain MC58 was cloned and expressed in the pRSET expression vector (section 2.19). The His-tagged α unit was purified on a Nickel gel and eluted with imidazole. The subunit was then used in gel shift assays using the *opc* promoter region

using the same promoter region from pC12R described in section 7.10, which demonstrated a clearly visible gel shift with an appropriately high mobility complex (Figure 7.13). This indicates that the α subunit of RNA polymerase has a high affinity direct interaction with the *opc* promoter sequence.

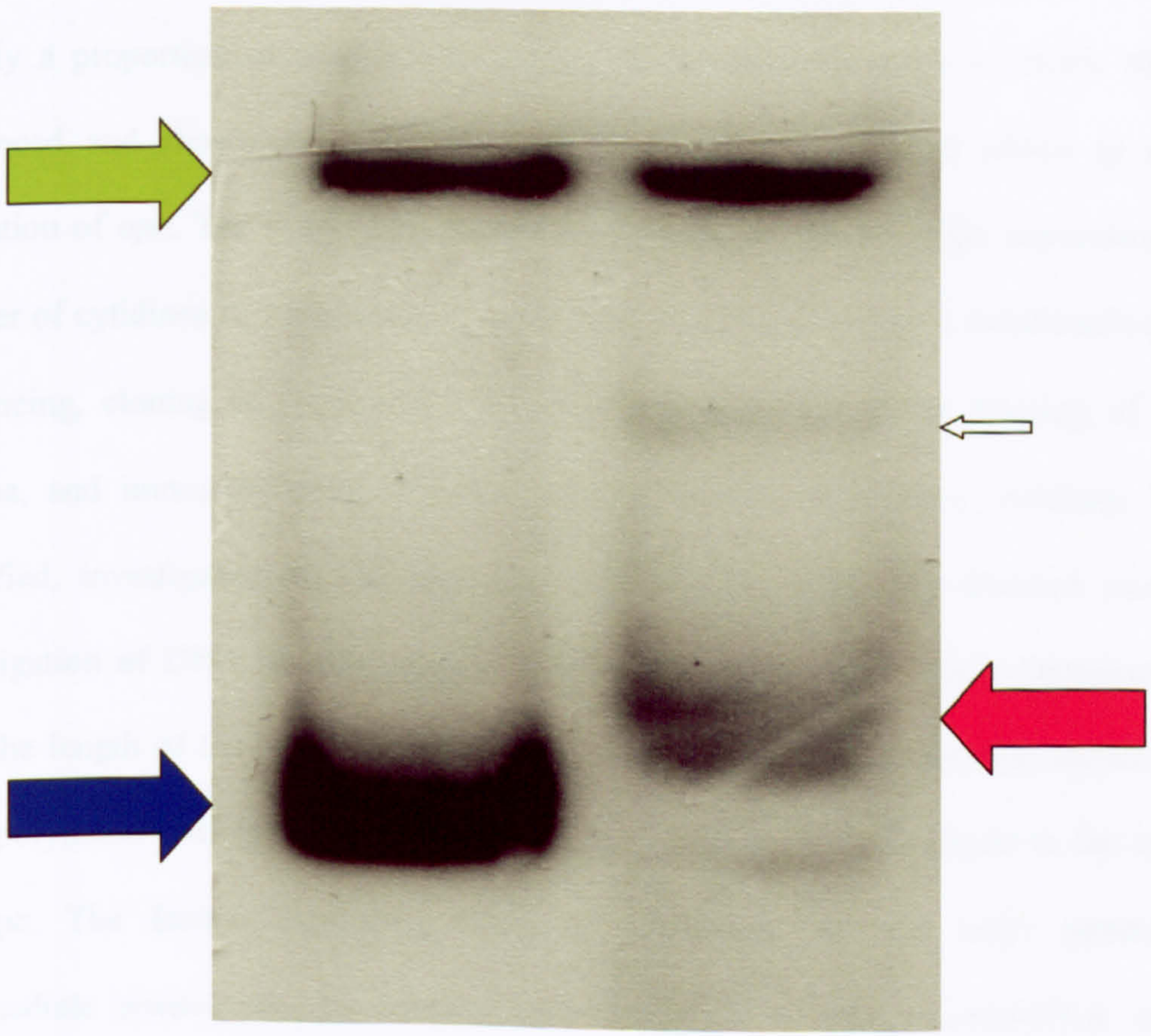
7.11 Mutagenesis of the putative IHF binding site in the *opc* promoter

There is a putative integration host factor (IHF) binding site in the promoter region of the *opc* gene (shown in figure 7.1). The presence of the high mobility complex in the EMSA assays is consistent with the specific binding of a small protein complex, such as IHF. The high concentration ammonium sulphate precipitation (as seen in the 80% ammonium sulphate precipitate shown in figure 7.10), in which protein forming the high mobility complex was still present after the RNA polymerase had been removed contained several low molecular weight proteins consistent with the presence of IHF. There were also a number of smaller proteins seen in the extract obtained by affinity purification using the *opc* promoter (figure 7.12). Taken together this suggested that the putative accessory protein might be IHF.

Using an overlapping PCR strategy (figure 2.2), site directed mutagenesis was used to alter the putative IHF site in the *opc* promoter (the altered bases are shown in figure 7.1). This was done in the already altered plasmid containing the replacement sequence equivalent to 13 Cs in the location of the homopolymeric tract that led to stable high-level Opc expression (pC13R), as described in section 2.18b. Three transformants with the altered IHF site were tested for Opc expression by colony immunoblotting using anti-Opc monoclonal antibody B306. The putative IHF binding site mutants did not express detectable Opc (data not shown). Subsequently the whole *opc* gene and the promoter were sequenced to exclude the possibility that a second mutation had occurred in the gene or promoter that would prevent expression of *opc*. The gene and promoter were found to be otherwise the same as the wild type gene.

Figure 7.13

EMSA showing the interaction between the α subunit of RNA polymerase and the *opc* promoter. The assay used a probe DNA consisting of 166 bp of *opc* promoter sequence. The probe was amplified with OpcPro and γ^{32}P labelled Opc16 from a plasmid with the replacement tract from pC13R, with a length equivalent to 13 Cs. The cloned, expressed and purified α subunit of RNA polymerase of *N. meningitidis* strain MC58 was added to the reaction shown in the left hand lane. The free probe band (blue arrow) and material retained in the well of the gel (green arrow) can be seen in the left hand lane. The promoter shifted by interaction with the α subunit of RNA polymerase can be seen in the right hand lane (red arrow). A minor artifactual band (small white arrow) is also seen in the right hand lane.



This provides support for the hypothesis that the IHF consensus sequence is an active binding site in the *opc* promoter and that expression of *opc* is dependent upon IHF binding to the promoter. This suggests a model in which IHF may control the expression of *opc* in a manner similar to that seen with the binding of RNA polymerase in *Esch. coli* when it is dependent upon the cyclic AMP receptor protein (CRP) (Savery *et al.*, 1995). In this situation the α-subunit of RNA polymerase is capable of binding to the promoter directly, however this is not able to mediate open complex formation, and hence transcription, in the absence of CRP. IHF binding to the *opc* promoter may act in a similar way. In *N. gonorrhoeae* IHF expression has been shown to decline as the cells enter stationary phase (Hill *et al.*, 1998). If the pattern of expression of IHF is similar in *N. meningitidis* this suggests that expression of Opc will be highest during periods of active cell growth.

7.12 Summary of the investigation of *opc*

The presence of *opc* was studied in serogroup B meningococci and was found to be present in only a proportion of strains. The gene and flanking sequence of strain MC58 was sequenced and shown to be an appropriate model organism in which to study the regulation of *opc*. The previously described correlation between Opc expression and the number of cytidines in the homopolymeric tract was confirmed by a combination of direct sequencing, cloning of promoters, restriction digestion and silver staining of promoter regions, and immunoblotting of Opc. After a source of *in vitro* variation had been identified, investigation of the promoter was pursued using site-directed mutants and investigation of DNA binding protein – promoter interactions. These demonstrated that both the length of the repeat (through a spacing effect) and the specific sequence of the homopolymeric tract (probably through a facing effect) each contribute to the expression of Opc. The former mediating ON-OFF switching and the latter generating the intermediate phenotypes. In addition the formation of two protein-DNA complexes between meningococcal DNA binding proteins and the promoter were demonstrated. The

larger complex formation, consistent with RNA polymerase, only occurs when the spacing sequence is consistent with expression, indicating that the binding of this complex is directly dependent upon the length of the homopolymeric tract. In addition, it has been demonstrated that the α subunit of RNA polymerase can interact directly with the *opc* promoter. The second, smaller, complex may be IHF and mutagenesis of the putative IHF binding site in the promoter leads to loss of expression of Opc. Taken together these experiments demonstrate that phase variation of *opc* is mediated directly by the homopolymeric tract by determining the interaction of RNA polymerase with the promoter. In addition, the expression of Opc is also independently regulated through an interaction involving the IHF consensus binding site that ends 42 bp upstream of the homopolymeric tract within the *opc* promoter.

Chapter 8

References

Abraham, J.M., Freitag, C.S. Clements, J.R. & Eisenstein, B.I. (1985). An invertible element of DNA controls phase variation of type 1 fimbriae of *Escherichia coli*. *Proc Natl Acad Sci USA* **82**: 5724-5727.

Abraham, J.M., Freitag, C.S., Gander, R.M., Clements, J.R., Thomas, V.L. & Eisenstein, B.L. (1986). Fimbrial phase variation and DNA arrangements in uropathogenic isolates of *Escherichia coli*. *Mol Biol Med* **3**: 495-508.

Achtman, M., Neibert, M., Crowe, B.A., Strittmatter, W., Kusecek, B., Weyse, E., Walsh, M.J., Slawig, B., Morelli, G., Moll, A. & Blake, M. (1988). Purification and characterization of eight class 5 outer membrane protein variants from a clone of *Neisseria meningitidis* serogroup A. *J Exp Med* **168**: 507-525.

Achtman, M, Wall, R.A., Bopp, M., Kusecek, B., Morelli, G., Saken, E. & Hassan-King, M. (1991a). Variation in class 5 protein expression by serogroup A meningococci during a meningitis epidemic. *J Infect Dis* **164**: 375-382.

Achtman, M., Morelli, G., Kusecek, M., Bopp, J. Wang, J. & Caugant, D. (1991b). Properties and epidemiology of 2 epidemic clones of serogroup A *Neisseria meningitidis* associated with African epidemics since 1980. In: Actman, M., Kohl, P., Marchal, C., *et al.*, Eds. Neisseria 1990. Walter de Gruyter & Co. Berlin, 1991: 5-10.

Aho, E.L. (1989). Molecular biology of Class 5 outer membrane proteins of *Neisseria meningitidis*. **7**: 249-253.

Aho, E.L., Dempsey, J.A., Hobbs, M.M., Klapper, D.G. & Cannon, J.G. (1991). Characterization of the opa (class 5) gene family of *Neisseria meningitidis*. *Mol Microbiol* **5**: 1429-1437.

Al-Mamun, A.A.M., Tominaga, A. & Enomoto, M. (1997). Cloning and characterisation of the regions III flagellar operons of the four Shigellar subgroups: Genetic defects that cause loss of flagella in *Shigella boydii* and *Shigella sonnei*. *J Bacteriol* **179**: 4493-4500.

Ala'Aldeen, D.A., Davies, H.A. & Borriello, S.P. (1994). Vaccine potential of meningococcal FrpB: studies on surface exposure and functional attributes of common epitopes. *Vaccine* **12**: 535-541.

Alm, R.A., Ling, L-S., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C,. Smith D.R., Noonan, B., Guild, B.C., deJonge, B.L. *et al.*, (1999). Genomic-sequences comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**: 176-180.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol **215**: 403-410.

Andrewes, F.W. (1922). Studies in group-agglulination. I. The *Salmonella* group and its antigenic structure. *J Path Bact* **25**: 505-521.

Apicella, M.A., Shero, M., Jarvis, G.A., Griffiss, J.M., Mandrell, R.E. & Schneider, H., (1987). Phenotypic variation in epitope expression of the *Neisseria gonorrhoeae* lipooligosaccharide. *Infect Immun* **55**: 1755-1761.

Apicella, M.A., Mandrell, R.E., Shero, M., Wilson, M.E., Groffiss, J.M., Brooks, G.F., Lammel, C., Breen, J.F. & Rice, P.E. (1990). Modification by sialic acid of *Neisseria gonorrhoeae* lipooligosaccharide epitope expression in human urethral exudates: an immunoelectron microscopic analysis. *J Infect Dis* **162**: 506-512.

Appelmelk, B.J., Negrini, R., Moran, A.P. & Kuipers, E.J. (1997). Molecular mimicry between *Helicobacter* pylori and the host. *Trends Microbiol* **5:** 70-73.

Appelmelk, B.J., Shiberu, B., Trinks, C., Tapsi, N., Zheng, P.Y., Verboom, T, Maaskant, J., Hokke, C.H., Schiphorst, W.E., Blanchard, D., Simoons-Smit, I.M., van den Eijnden, D.H. & Vandenbroucke-Grauls, C.M. (1998). Phase variation in *Helicobacter pylori* lipopolysaccharide. *Infect Immun* **66**: 70-76.

Appelmelk, B.J., Martin, S.L., Monteiro, M.A., Clayton, C.A., McColm, A.A., Zheng, P., Verboom, T., Maaskant, J.J., van den Eijnden, D.H., Hokke, C.H., Perry, M.B., Vandenbroucke-Grauls, C.M. & Kusters, J.G. (1999). Phase variation in *Helicobacter pylori* lipopolysaccharide due to changes in the lengths of poly(C) tracts in alpha3-fucosyltransferase genes. *Infect Immun* **67**: 5361-5366.

Aricó, B., Miller, J.F., Roy, C., Stibitz, S., Monack, D., Falkow, S., Gross, R. & Rappuoli, R. (1989). Sequences required for expression of *Bordetella pertussis* virulence factors share homology with prokaryotic signal transduction proteins. *Proc Natl Acad Sci U S A* **86**: 6671-6675.

Armitage, P.J. (1952). The statistical theory of bacterial populations subject to mutation. *J Roy Statistical Soc* **B14**: 1-40.

Armitage, P.J. (1953). Statistical concepts in the theory of bacterial mutation. *J Hygiene* **51**: 162-184.

Arora, S.K., Richings, B.W., Almira, E.C., Lory, S. & Ramphal, R. (1998). The *Pseudomonas aeruginosa* flagellar cap protein, FliD, is responsible for mucin adhesion. *Infect Immun* **66**: 1000-1007.

Ashworth, L.A., Irons, L.I. & Dowsett, A.B. (1982). Antigenic relationship between serotype-specific agglutinogen and fimbriae of *Bordetella pertussis*. *Infect Immun* **37**: 1278-1281.

Asteris, G. & Sarkar, S. (1996). Bayesian procedures for the estimation of mutation rates from fluctuation experiments. *Genetics* **142**: 313-326.

Athamna, A., Rosengarten, R., Levinson, S., Kahane, I. & Yogev, D. (1997). Adherence of *Mycoplasma gallisepticum* involves variable surface membrane proteins. *Infect Immun* **65**: 2468-2471.

Auble, D.T., Allen, T.L. & deHaseth, P.L. (1986). Promoter recognition by *Escherichia coli* RNA polymersase. Effects of substitutions in the spacer DNA separating the –10 and –35 regions. *J Biol Chem* **261**: 11202-11206.

Auble, D.T. & deHaseth, P.L. (1988). Promoter recognition by *Escherichia coli* RNA polymerase. Influence of DNA structure in the spacer separating the –10 and –35 regions. *J Mol Biol* **202**: 471-482.

Ayers, D.G., Auble, D.T. & deHaseth, P.L. (1989). Promoter recognition by *Escherichia co*li RNA polymerase. Role of the spacer DNA in functional complex formation. *J Mol Biol* **207**: 749-756.

Azmi, F.H., Lucas, A.H., Spiegelberg, H.L. & Granoff, D.M. (1995). Human immunoglobulin M paraproteins cross-reactive with *Neisseria meningitidis* group B polysaccharide and fetal brain. *Infect Immun* **63**: 1906-1913.

Baehr, W., Gotschlich, E.C. & Hitchcock, P.J. (1989). The virulence-associated gonococcal H.8 encodes 14 tandemly repeated pentapeptides. *Mol Microbiol* **3**: 49-55.

Bairoch, A. & Apweller, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res* **25:** 31-36.

Banemann, A. & Gross, R. (1997). Phase variation affects long term survival of *Bordetella bronchiseptica* in professional phagocytes. *Infect Immun* **65**: 3469-3473.

Banemann, A., Deppisch, H. & Gross, R. (1998). The lipopolysaccharide of *Bordetella bronchiseptica* acts as a protective shield against antimicrobial peptides. *Infect Immun* **66**: 5607-5612.

Banerjee, A., Wang, R., Uljon, S.N., Rice, P.A., Gotschlich, E.C. & Stein, D.C. (1998). Identification of the gene (*lgtG*) encoding the lipooligosaccharide ß chain synthesizing glucosyl transferase from *Neisseria gonorhoeae. Proc Natl Acad Sci USA* **95**: 10872-10877.

Bar-Shavit, Z., Ofek, I., Goldman, R., Mirelman, D. & Sharon, N. (1977). Mannose residues on phagocytes as receptors for the attachment of *Escherichia coli* and *Salmonella typhi. Biochem Biophys Res Commun* **78**: 455-460.

Barbour, A.G., Tessier, S.L. & Stoenner, H.G. (1982). Variable major proteins of *Borrelia hermsii. J Exp Med* **156**: 1312-1324.

Barbour, A.G., Barrera, O. & Judd, R.C. (1983). Structural analysis of the variable major proteins of *Borrelia hermsii. J Exp Med* **158**: 2127-2140.

Barlow, A.K., Heckels, J.E. & Clarke, I.N. (1989). The class 1 outer membrane protein of *Neisseria meningitidis* gene sequence and structural and immunological similarities to gonococcal porins. *Mol Microbiol* **3**: 131-139.

Bartlett, D.H., Wright, M.E. & Silverman, M. (1988). Variable expression of extracellular polysaccharide in the marine bacterium *Pseudomonas atlantica* is controlled by genome rearrangement. *Proc Natl Acad Sci USA* **85**: 3923-3927.

Baseman, J.B., Morrison-Plummer, J., Drouillard, D., Puleo-Scheppke, B., Tryon, V.V. & Holt, S.C. (1987). Identification of a 32-kilodalton protein of *Mycoplasma pneumoniae* associated with hemadsorption. *Isr J Med Sci* **23**: 474-479.

Bauer, F.J., Rudel, T., Stein, M. & Meyer, T.F. (1999). Mutagenesis of the *Neisseria gonorrhoeae* porin reduces invasion in epithelial cells and enhances phagocyte responsiveness. *Mol Microbiol* **31**: 903-913.

Baumler, A.J., Kusters, J.G., Stojiljkovic, I. & Heffron, F. (1994). *Salmonella typhimurium* loci involved in survival within macrophages. *Infect Immun* **62**: 1623-1630.

Behrens, A., Heller, M., Kirchhoff, H., Yogev, D. & Rosengarten, R. (1994). A family of phase- and size-variant membrane surface lipoprotein antigens (Vsps) of *Mycoplasma bovis*. *Infect Immun* **62**: 5075-5084.

Behrens, A., Poumarat, F., Le Grand, D., Heller, M. & Rosengarten, R. (1996). A newly identified immunodominant membrane protein (pMB67) involved in *Mycoplasma bovis* surface antigenic variation. *Microbiology* **142**: 2463-2470.

Belland, R.J., Morrison, S.G., van der Lay, P. & Swanson, J. (1989). Expression and phase variation of gonococcal P.II genes in *Escherichia coli* involves ribosomal frameshifting and slipped-strand mispairing. *Mol Microbiol* **3**: 777-786.

Belland, R.J., Chen, T., Swanson, J. & Fischer, S.H. (1992). Human neutrophil response to recombinant neisserial Opa proteins. *Mol Microbiol* **6**: 1729-1737.

Belland, R.J., Morrison, S.G. & Hogan, D. (1996). A phase-variable type III restriction-modification system in *Neisseria gonorrhoeae*. In: Abstracts of the Tenth International Pathogenic Neisseria Conference: poster 117, p. 360-361.

Belland, R.J., Morrison, S.G., Carlson, J.H. & Hogan, D.M. (1997). Promoter strength influences phase variation of neisserial *opa* genes. *Mol Microbiol* **23**: 123-135.

Bessen, D. & Gotschlich, E.C. (1986). Interactions of gonococci with HeLa cells: attachment, detachment, replication, penetration, and the role of protein II. *Infect Immun* **54**: 154-160.

Beucher, M. & Sparling, P.F. (1995). Cloning, sequencing, and characterisation of the gene encoding FrpB, a major iron-regulated, outer membrane protein of *Neisseria gonorrhoeae*. *J Bacteriol* **177**: 2041-2049.

Bhat, K.S., Gibbs, C.P., Barrera, O., Morrison, S.G., Jahnig, F., Stern, A., Kupsch, E.-M., Meyer, T.F. & Swanson, J. (1991). The opacity proteins of *Neisseria gonorrhoeae* strain MS11 are encoded by a family of 11 complete genes. *Mol Microbiol* **5**: 1889-1901. **also see erratum**: *Mol Microbiol* **6**: 1073-1076.

Bhugra, B. & Dybvig, K. (1992). High frequency rearrangements in the chromosome of *Mycoplasma pulmonis* correlate with phenotypic switching. *Mol Microbiol* **6**: 1149-1154.

Bhugra, B, Voelker, L.L, Zou, N., Yu, H. & Dybvig, K. (1995). Mechanism of antigenic variation in *Mycoplasma pulmonis*: interwoven, site-specific DNA inversions. *Mol Microbiol* **18**: 703-714.

Biswas, G.D. & Sparling, P.F. (1995). Characterisation of *lbpA*, the structural gene for a lactoferrin receptor in *Neisseria gonorrhoeae*. *Infect Immun* **63**: 2958-2967.

Blake, M.S., Blake, C,M., Apicella, M.A. & Mandrell, R.E. (1995). Gonococcal opacity: lectin-like interactions between Opa proteins and lipopolysaccharide. *Infect Immun* **63**: 1434-1439.

Blomfield, I.C., Calie, P.J., Eberhardt, K.J., McClain, M.S. & Eisenstein, B.I. (1993). Lrp stimulates the phase variation of type 1 fimbriation in Escherichia coli K12. *J Bacteriol* **175**: 27-36.

Blyn, L.B., Braaten, B.A. & Low, D.A. (1990). Regulation of *pap* pilin phase variation by a mechanism involving differnetial *dam* methylation states. *EMBO J* **9**: 4045-4054.

Braaten, B.A., Blyn, L.B., Skinner, B.S. & Low, D.A. (1991). Evidence for a methylation-blocking factor (*mbf*) locus involved in *pap* pilus expression and phase variation in *Escherichia coli*. *J Bacteriol* **173**: 1789-1800.

Braaten, B.A., Nou, X., Kaltenbach, L.S. & Low, D.A. (1994). Methylation patterns in *pap* regulatory DNA controlling the pyelonephritis-associated pili phase variation in *E. coli*. *Proc Natl Acad Sci USA* **89**: 4250-4254.

Brennan, M.J., David, J.L., Kenimer, J.G. & Manclark, C.R. (1988). Lectin-like binding of pertussis toxin to a 165-kilodalton Chinese hamster ovary cell glycoprotein. *J Biol Chem* **263**: 4895-4899.

Brinton, C.C. (1959). Non-flagellar appendages of bacteria. *Nature* **183**: 782-786.

Brinton, C.C., Jr., Carter, M.J., Berber, D.B., Kar, S., Kramarik, J.A., To, A.C.-C., To, S.C.-M., and Wood, S.W. (1989). Design and development of pilus vaccines for *Haemophilus influenzae* diseases. *Pediatr Infect Dis J* **8** (Suppl): S54-S61.

Boren, T., Falk, P., Roth, K.A., Larson, G. and Normark, S. (1993). Attachment of *Helicobacter pylori* to humangastric epithelium mediated by blood group antigens. *Science* **262**: 1892-1895.

Bucci, C., Lavitola, A., Salvatore, P., Giudice, L.D., Masardo, D.R., Bruni, C.B. & Alifano. (1999). Hypermutation in pathogenic bacteria: frequent phase variation in meningococci is a phenotypic trait of a specialized mutator biotype. *Mol Cell* **3**: 435-445.

Bukhari, A.I. & Ambrosio, L. (1978). The invertible segment of bacteriophage Mu DNA determines the absorption properties of Mu particles. *Nature* **271**: 575-577.

Bull, J.J. & Pease, C.M. (1995). Why is the polymerase chain reaction resistant to in vitro evolution? *J Mol Evol* **41**: 1160-1164.

Bunting, M.I. (1940). The production of stable populations of color variants of *Serratia marcescens* no. 274 in rapidly growing cultures. *J Bacteriol* **40**: 69-81.

Burgess, R.R. (1969). A new method for the large scale purification of *Escherichia coli* deoxyribonucleic acid-dependent ribonucleic acid polymerase. *J Biol Chem* **244**: 6160-6167.

Burgess, R.R. & Jendrisak, J.J. (1975). A procedure for the rapid, large-scale purification of *Escherichia coli* DNA-dependent RNA polymerase involving polymin P precipitation and DNA-cellulose chromatography. *Biochemistry* **14**: 4634-3628.

Burch, C.L., Danaher, R.J. & Stein, D. (1997). Antigenic variation in *Neisseria gonorrhoeae*: production of multiple lipopolysaccharides. *J Bacteriol* **179**: 982-986.

Busby, S. & Ebright, R.H. (1994). Promoter structure, promoter recognition, and transcription activation in prokaryotes. *Cell* **79**: 743-746.

Cairns, J., Overbaugh, J. & Miller, S. (1988). The origin of mutants. *Nature* **335**: 142-145.

Cannon, J.G., Black, W.J., Nachamkin, I. & Stewart, P.W. (1984). Monoclonal antibody that recognizes an outer membrane antigen common to the pathogenic *Neisseria* species but not to most nonpathogenic *Neisseria* species. *Infect Immun* **43**: 994-999.

Carbonetti, N.H., Khelef, N., Guiso, N. & Gross, R. (1993). A phase variant of *Bordetella pertussis* with a mutation in a new locus involved in the regulation of pertussis toxin and adenylate cyclase toxin expression. *J Bacteriol* **175**: 6679-6688.

Caroff, M., Chaby, R., Karibian, D., Perry, J., Deprun, C. & Szabo, L. (1990). Variations in the carbohydrate regions of *Bordetella pertussis* lipopolysaccharides: eletrophoretic, serological, and structural features. *J Bacteriol* **172**: 1121-1128.

Carroll, P.A., Tashima, K.T., Rogers, M.B., DiRita, V.J. & Calderwood, S.B. (1997). Phase variation in *tcpH* modulates expression of the ToxR regulon in *Vibrio cholerae*. *Mol Microbiol* **25**: 1099-1111.

Cartwright, K.A.V., Stuart, J.M., Jones, D.M. & Noah, N.D. (1987) The Stonehouse survey: nasopharyngeal carriage of meningococci and *Neisseria lactamica*. *Epidem Infect* **99**: 591-601.

Cattaneo, L.A., Reed, G.W., Haase, D.H., Wills, M.J. & Edwards, K.M. (1996). The seroepidemiology of *Bordetella pertussis* infections: A study of persons ages 1 – 65 years. *J Infect Dis* **173**: 1256-1259.

Chamberlin, M., Kingston, R., Gilman, M., Wiggs, J. & deVera, A. (1983). Isolation of bacterial and bacteriophage RNA polymerases and their use in synthesis of RNA *in vitro*. *Methods Enzymol* **101**: 540-568.

Chen, C-J., Sparling, P.F., Lewis, L.A., Dyer, D.W. & Elkins, C. (1996). Identification and purification of a hemoglobin-binding outer membrane protein from *Neisseria gonorhoeae*. *Infect Immun* **64**: 5008-5014.

Chen, C.J., Elkins, C. & Sparling, P.F. (1998) Phase variation of hemoglobin utilization in *Neisseria gonorrhoeae*. *Infect Immun* **66**: 987-993.

Citti, C. & Wise, K.S. (1995) *Mycoplasma hyorhinis vlp* gene transcription: critical role in phase variation and expression of surface lipoproteins. *Mol Microbiol* **18**: 649-660.

Citti, C., Kim, M.F. & Wise, K.S. (1997). Elongated versions of Vlp surface lipoproteins protect *Mycoplasma hyrhinis* escape variants from growth-inhibiting host antibodies. *Infect Immun* **65**: 1773-1785.

Coffey, E.M. & Eveland, W.C. (1967). Experimental relapsing fever initiated by *Borrelia hermsii*. II. Sequential appearance of major serotypes in the rat. *J Infect Dis* **177**: 29-34.

Collins, R., Achtman, M., Ford, R., Bullough, P. & Derrick, J. (1999). Projection structure of reconstituted Opc outer membrane protein from *Neisseria meningitidis*. *Mol Microbiol* **32**: 217-219.

Connell, T.D., Black, W.J., Kawula, T.H., Barritt, D.S., Dempsey, J.A., Kverneland, K. Jr., Stephenson, A., Schepart, B.S., Murphy, G.L. & Cannon, J.G. (1988). Recombination among protein II genes of *Neisseria gonorrhoeae* generates new coding sequences and increases structural variability in the protein II family. *Mol Microbiol* **2**: 227-236.

Connel, T.D., Shaffer, D. & Cannon, J.G. (1990). Characterization of the repertoire of hypervariable regions in the Protein II (*opa*) gene family of *Neisseria gonorrhoeae*. *Mol Microbiol* **4**: 439-449.

Connor, E.M. & Loeb, M.R. (1983). A hemadsorption method for detection of colonies of *Haemophilus influenzae* type b expressing fimbriae. *J Infect Dis* **148**: 855-859.

Cope, L.D., Yogev, R., Mertsola, J., Argyle, J.C., McCraken Jr, G.H. & Hansen, E.J. (1990). Effect of mutations in lipopolysaccharide biosynthesis genes on virulence of *Haemophilus influenzae* type b. *Infect Immun* **58**: 2343-2351.

Cope, L.D., Yogev, R., Mertsola, J., Latimer, J.L., Hanson, M.S., McCraken, G.H., Jr. & Hansen, E.J. (1991). Molecular cloning of a gene involved in lipopolysaccharide biosynthesis and virulence expression by *Haemophilus influenzae* type B. *Mol Microbiol* **5**: 1113-1124.

Cope, L.D., Yogev, R., Muller-Eberhard, U. & Hansen, E.J. (1995). A gene cluster involved in the utilization of both free heme and heme:hemopexin by *Haemophilus influenzae* type b. *J Bacteriol* **177**: 2644-2653.

Cotter, P.A., Yuk, M.H., Mattoo, S., Akerley, B.J., Boschwitz, J., Relman, D.A. & Miller, J.F. (1998). Filamentous hemagglutinin of *Bordetella bronchiseptica* is required for efficient establishment of tracheal colonization. *Infect Immun* **66**: 5921-5929.

Cox, D.R. & Miller, H.D. (1965). The Theory of Stochastic Processes. Chapman & Hall.

Cox, D.L., Chang, P., McDowall, A.W. & Radolf, J.D. (1992). The outer membrane, not a coat of host proteins, limits antigenicity of virulent *Treponema pallidum*. *Infect Immun* **60**, 1076-1083.

Crowe, B.A., Wall, R.A., Kusecek, B., Neumann, B., Olyhoek, T., Abdillahi, H., Hassan-King, M., Greenwood, B.M., Poolman, J.T. & Achtman, M. (1989). Clonal and variable properties of *Neisseria meningitidis* isolated from cases and carriers during and after an epidemic in the Gambia, West Africa. *J Infect Dis* **159**: 686-700.

Dallo, S.F. & Baseman, J.B. (1991). Adhesin gene of *Mycoplasma genitalium* exists as multiple copies. *Microb Pathog* **10**: 475-480.

Danaher, R.J., Levin, J.C., Arking, d., Burch, C.L., Sandlin, R. & Stein, D.C. (1995). Genetic basis of *Neisseria gonorrhoeae* lipooligosaccharide antigenic variation. *J Bacteriol* **177**: 7275-7279.

Davis, D.J., Pittman, M. & Griffitts, J.J. (1950). Hemagglutination by the Koch-Weeks bacillus (*Hemophilus aegyptius*). *J Bacteriol* **59**: 427-431.

Davison, B.L., Leighton, T. & Rabinowitz, J.C. (1979). Purification of *Bacillus subtilis* RNA polymerase with heparin-agarose. *J Biol Chem* **254**: 9220-9226.

De Bolle, X., Bayliss, C.D., van de Ven, T., Saunders, N.J., Hood, D.W. & E. Richard Moxon. (1999). The length of a tetranucleotide repeat tract in *Haemophilus influenzae* determines the phase variation rate of a gene with homology to type III DNA methyltransferase. *Mol Microbiol* **35**: 211-222.

de Cossio, M.E.F., Ohlin, M., Llano, M., Selander, B., Cruz, S., del Valle, J. & Borrebaeck, C.A.K. (1992). Human monoclonal antibodies against an epitope on the class 5c outer membrane protein common to many pathogenic strains of *Neisseria meningitidis*. *J Infect Dis* **166**: 1322-1328.

de Gier, J-W.L., Schepper, M., Reijnders, W.N.M., van Dyck, S.J., Slotboom, D.J., Warne, A., Saraste, M., Krab, K., Finel, M., Stouthamer, A.H., van Spanning, R.J.M. & van der Oost, J. (1996). Structural and functional analysis of the aa_3-type and cbb_3-type cytochrome *c* oxidases of *Paracoccus denitrificans* reveals significant differences in proton-pump design. *Mol Microbiol* **20**: 1247-1260.

Dekker, N.P., Lammel, C.J., Mandrell, R.E. & Brooks, G.F. (1990). Opa (protein II) influences gonococcal organisation in colonies, surface appearance, size and attachment to human fallopian tube tissues. *Microb Pathog* **9**: 19-31.

de la Paz, H., Cooke, S.J. & Heckels, J.E. (1995). Effect of sialylation of lipopolysaccharide of *Neisseria gonorrhoeae* on recognition and complement-mediated killing by monoclonal antibodies directed against different outer-membrane antigens. *Microbiology* **141**: 913-920.

De Magistris, M.T., Romano, M., Nuti, S., Rappouli, R. & Tagliabue, A. (1988). Dissecting human T cell responses against *Bordetella* species. *J Exp Med* **168**: 1351-1362.

Demerec, M. (1945). Production of staphylococcus strains resistant to various concentrations of penicillin. *Proc Natl Acad Sci Wash* **31**: 16-24.

Dempsey, J.-A.F., Litaker, W., Madhure, A., Snodgrass, T.L. & Cannon, J.G. (1991). Physical map of the chromosome of *Neisseria gonorrhoeae* FA1090 with locations of genetic markers, including *opa* and *pil* genes. *J Bacteriol* **173**: 5476-5486.

Deretic, V., Schurr, M.J. & Yu Hongwei. (1995). *Pseudomonas aeruginosa*, mucoidy and the chronic infection phenotype in cystic fibrosis. *Trends Microbiol* **3**: 351-356.

Deville, J.G., Cherry, J.D., Christenson, P.D., Pineda, E., Leach, C.T., Kuhls, T.L. & Viker, S. (1995). Frequency of unrecognized *Bordetella pertussis* infections in adults. *Clin Infect Dis* **21**: 639-642.

DeVoe, I.W. (1982). The meningococcus and mechanisms of pathogenicity. *Microbiol Rev* **46**: 162-190.

de Vries, F.P., van der Ende, A., van Putten, J.P.M. & Dankert, J. (1996). Invasion of primary nasopharyngeal epithelial cells by *Neisseria meningitidis* is controlled by phase variation of multiple surface antigens. *Infect Immun* **64**: 2998-3006.

de Vries, F.P., Cole, R., Dankert, J., Frosch, M. & van Putten, J.P.M. (1998). *Neisseria meningitidis* producing the Opc adhesin binds epithelial cell proteoglycan receptors. *Mol Microbiol* **27**: 1203-1212.

DiRita, V.J., Parsot, C., Jander, G. & Mekalanos, J.J. (1991). Regulatory cascade controls virulence in *Vibrio cholera*. *Proc Natl Acad Sci USA* **88**: 5403-5407.

Dorman, C.J. & Higgins, C.F. (1987). Fimbrial phase variation in *Escherichia coli*: dependence on integration host factor and homologies with other site-specific recombinases. *J Bacteriol* **169**: 3840-3843.

Drake, J.W. (1991). A constant rate of spontaneous mutation in DNA-based microbes. *Proc Natl Acad Sci USA* **88**: 7160-7164.

Duensing, T.D. & van Putten, J.P.M. (1997). Vitronectin mediates internalisation of *Neisseria gonorrhoeae* by Chinese hamster ovary cells. *Infect Immun* **65**: 964-970.

Durbin, R. & Mieg, J.T. (1991). A C. elegans Database. Documentation, code and data available from anonymous FTP servers at: lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov.

Dybvig, K. & Yu, H. (1994) Regulation of a restriction and modification system via DNA inversion. *Mol Microbiol* **12**: 547-560.

Dyer, D.W., West, E.P., McKenna, W., Thompson, S.A. & Sparling, P.F. (1988). A pleiotrophic iron-uptake mutant of *Neisseria meningitidis* lacks a 70-kilodalton iron-regulated protein. *Infect Immun* **56**: 977-983.

Ebright, R.H. & Busby, S. (1995). The *Escherichia coli* RNA polymerase alpha subunit: structure and function. *Curr Opin Genet Dev* **5**: 197-203.

Eisenstein, B.I. (1981). Phase variation of type 1 fimbriae in *Escherichia coli* is under transcriptional control. *Science* **214**: 337-339.

Eisenstein, B.I., Sweet, D.S., Vaughn, V. & Friedman, D.I. (1987). Integration host factor is required for the DNA inversion that controls phase variation in *Escherichia coli*. *Proc Natl Acad Sci USA* **84**: 6506-6510.

Ellinger, T., Behnke, D., Knaus, R. & Bujard, J.D. (1994). Context-dependent effects of upstream A-tracts. *J Mol Biol* **239**: 466-475.

Emmert, D.B., Stoehr, P.J., Stoesser, G. & Cameron, G.N. (1994). The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Res* **22**: 3445-3449.

Englard, S. & Seifter, S. (1990). Precipitation techniques. *Methods Enzymol* **182**: 285-300.

Enomoto, M. & Stocker, B.A.D. (1975). Integration, at hag or elsewhere, of H2 (phase-2 flagellin) genes transduced from Salmonella to Escherichia coli. *Genetics* **81**: 595-614.

Etzold, T. & Argos, P. (1993). SRS an indexing and retrieval tool for flat data libraries. *Comput Appl Biosci* **9**: 49-57.

Facius, D. & Meyer, T.F. (1993). A novel determinant (*comA*) essential for natural transformation competence in *Neisseria gonorrhoeae* and the effect of a *comA* defect on pilin variation. *Mol Microbiol* **10**: 699-712.

Farley, M.M., Stephens, D.S., Kaplan, S.L. & Mason, E.O., Jr. (1990). Pilus- and non-pilus-mediated interactions of *Haemophilus influenzae* type b with human erythrocytes and human nasopharyngeal mucosa. *J Infect Dis* **161**: 274-280.

Fearon, D.T. (1978). Regulation by membrane sialic acid of ß1H-dependent decay-dissociation of amplification C3 convertase of the alternative complement pathway. *Proc Natl Acad Sci USA* **75**: 1971-1975.

Fenno, J.C., Wong, G.W.K., Hannam, P.M., Muller, K-H., Leung, W.K. & McBride, B.C. (1997). Conservation of *msp*, the gene encoding the major outer membrane protein of oral *Treponema* spp. *J Bacteriol* **179**, 1082-1089.

Finlay, B.B. & Falkow, S. (1989). Common themes in microbial pathogenicity. *Microbiol Rev* **53**: 210-230.

Finne, J., Leinonen, M. & Makela, P.H. (1983). Anitgenic similarities between brain components and bacteria causing meningitis. Implications for vaccine development and pathogenesis. *Lancet* **2**: 355-357.

Finne, J., Bitter-Suermann, D., Goridis, C. & Finne, U. (1987). An IgG monoclonal antibody to group B meningococci cross-reacts with developmentally regulated polysialic acid units of glycoproteins in neural and extraneural tissues. *J Immunol* **138**: 4402-4407.

Fischer, S.H. & Rest, R.F. (1988). Gonococci possessing only certain P.II outer membrane proteins interact with human neutrophils. *Infect Immun* **56**: 1574-1579.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F. & Kerlavage, A.R. *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496-498, 507-512.

Flynn, J.L. & Ohman, D.E. (1988). Cloning of genes from mucoid *Pseudomonas aeruginosa* which control spontaneous conversion to the alginate production phenotype. *J. Bacteriol* **170**: 1452-1460.

Foster. P.L. (1999). Are adaptive mutations due to a decline in mismatch repair? The evidence is lacking. *Mutation Res* **436**: 179-184.

Frasch, C.E. & Mocca, L.F. (1978). Heat-modifiable outer membrane proteins of *Neisseria meningitidis* and their organisation within the membrane. *J Bacteriol* **178**: 1127-1134.

Fraser, C.M., Norris, S.J., Weinstock, G.M., White, O., Sutton, G.G. et al (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375-388.

Fujita, H., Yamaguchi, S. & Iino, T. (1973). Studies on H-O variants in *Salmonella* in relation to phase variation. *J Gen Microbiol* **76**: 127-134.

Gally, D.L., Bogan, J.A., Eisenstein, B.I. & Blomfield, I.C. (1993). Environmental regulation of the *fim* switch controlling type 1 fimbrial phase variation in *Escherichia coli* K-12: effects of temperature and media. *J Bacteriol* **175**: 6186-6193.

Gally, D.L., Rucker, T.J. & Blomfield, I.C. (1994). The leucine-responsive protein binds to the *fim* switch to control phase variation of type 1 fimbrial expression in *Escherichia coli*. *J Bacteriol* **176**: 5665-5672.

Gally, D.L., Leathart, J. & Blomfield, I. (1996). Interaction of FimB and FimE with the *fim* switch that controls the phase variation of type 1 fimbriae in *Escherichia coli* K12. *Mol Microbiol* **21**: 725-738.

Gaston, K., Bell, A., Kolb, A., Buc, H. & Busby, S. (1990). Stringent spacing requirements for transcription activation by CRP. *Cell* **62**: 733-743.

George, D.G., Barker, W.C., Mewes, H.W., Pfeiffer, F. & Tsugita, A. (1994). The PIR international protein sequence database. Nucleic Acids Res **22**: 3569-3573.

Giardina, P.C., Apicella, M.A., Gibson, B. & Preston, A. (1999). Antigenic mimicry in *Neisseria* species. In: Brade, H., Opal, S.M., Vogel, S.N. & Morrison, D.C. Endotoxin in Health and Disease. Marcel Dekker, New York.

Gibbs, C.P., Reimann, B.Y., Schultz, E., Kaufmann, A., Haas, R. & Meyer, T.F. (1989). Reassortment of pilin genes in *Neisseria gonorrhoeae* occurs by two distinct mechanisms. *Nature* **338**: 651-652.

Gill, M.J., McQuillen, D.P., van Putten, J.P.M., Wetzler, L.M., Bramley, J., Crooke, H., Parsons, N.J., Cole, J.A. & Smith, H. (1996). Functional characterization of a sialyltranferase-deficient mutant of *Neisseria gonorrhoeae*. *Infect Immun* **64**: 3374-3378.

Gilsdorf, J.R. & Ferrieri, P. (1986). Susceptibility of phenotypic variants of *Haemophilus influenzae* type b to serum bactericidal activity: relationship to surface lipopolysaccharide. *J Infect Dis* **153**: 223-231.

Gilsdorf, J.R., McCrea, K.W. & Marrs, C.F. (1997). Role of pili in *Haemophilus influenzae* adherence and colonisation. *Infect Immun* **65**: 2997-3002.

Givaudan, A., Lanois, A. & Boemare, N. (1996). Cloning and nucleotide sequence of a flagellin encoding genetic locus from *Xenorhabdus nematophilus*: phase variation leads to differential transcription of two flagellar genes (fliCD). *Gene* **183**, 243-253.

Gorby, G., Simon, D. & Rest, R,F. (1994). *Escherichia coli* that express *Neisseria gonorrhoeae* opacity-associated proteins attach to and invade human fallopian tube epithelium. *Ann N Y Acad Sci* **730**: 286-289.

Gotschlich, E.C. (1994). Genetic locus for the biosynthesis of the variable portion of *Neisseria gonorrhoeae* lipooligosaccharide. *J Exp Med* **180**: 2181-2190.

Gross, C., Engbaek, F., Flamming, T. & Burgess, R. (1976). Rapid micromethod for the purification of *Escherichia coli* ribonucleic acid polymerase and the preparation of bacterial extracts active in ribonucleic acid synthesis. *J Bacteriol* **128**: 382-389.

Gross, R. & Rappuoli, R. (1989). Pertussis toxin promoter sequences involved in modulation. *J Bacteriol* **171**: 4026-4030.

Guerina, N.G., Langermann, S., Clegg, H.W., Kessler, T.W., Goldmann, B.A. & Gilsdorf, J.R. (1982). Adherence of piliated *Haemophilus influenzae* type b to human oropharyngeal cells. *J Infect Dis* **146**: 564.

Guerry, P., Logan, S.M. & Trust, T.J. (1988). Genomic rearrangements associated with antigenic variation in *Campylobacter coli*. *J Bacteriol* **170**: 316-319.

Haas, R. & Meyer, T.F. (1986). The repertoire of silent pilus genes in *Neisseria gonorrhoeae*: evidence for gene conversion. *Cell* **44**: 107-115.

Haas, R., Schwarz, H. & Meyer, T.F. (1987). Release of soluble pilin coupled with gene conversion in *Neisseria gonorrhoeae*. *Proc Natl Acad Sci USA* **84**: 99079-9083.

Haines, K.A., Reibman, J., Tang, X.Y., Blake, M. & Weissmann, G. (1991). Effects of protein I of *Neisseria gonorrhoeae* on neutrophil activation: generation of diacylglycerol from phosphatidylcholine via a specific phospholipase C is associated with exocytosis. *J Cell Biol* **114**: 433-442.

Hammerschmidt, S., Birkholtz, C., Zahringer, U., Robertson, B.D., van Putten, J., Ebeling, O. & Frosch, M. (1994). Contribution of genes from the capsule gene cluster complex (*cps*) to lipopolysaccharide biosynthesis and serum resistance in *Neisseria meningitidis*. *Mol Microbiol* **11**: 885-896.

Hammerschmidt, S., Hilse, R., van Putten, JPM., Gerardy-Schahn, R., Unkmeir, A. & Frosch, M. (1996a). Modulation of cell surface sialic acid expression in *Neisseria meningitidis* via a transposable genetic element. *EMBO J* **15**: 192-198.

Hammerschmidt, S., Muller, A., Sillmann, H., Muhlenhoff, M., Borrow, R., Fox, A., van Putten, J., Zillinger, W.D., Gerardy-Schahn, R. & Frosch, M. (1996b). Capsule phase variation in *Neisseria meningitidis* serogroup B by slipped-strand mispairing in the polysialyltransferase gene (*siaD*): correlation with bacterial invasion and the outbreak of meningococcal disease. *Mol Microbiol* **20**: 1211-1220.

Hancock, J.M. & Armstrong, J.S. (1994). SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Comput Appl Biosci* **10:** 67-70.

Haneberg, B., Dalseg, R., Oftung, F., Wedege, E., Hoiby, E.A., Haugen, I.L., Holst, J., Andersen, S.R., Aase, A., Meyer Naess, L., Michaelsen, T.E., Namork, E. & Haaheim. L.R. (1998). Towards a nasal vaccine against meningococcal disease, and prospects for its use as a mucosal adjuvant. *Dev Biol Stand* **92**: 127-133.

Harnett, W. & Harnett, M.M. (1999). Phosphorylcholine: friend or foe of the immune system? *Immmunol Today* **20**: 125-129.

Harris, L.A., Logan, S.M., Guerry, P. & Trust, T.J. (1987). Antigenic variation of *Campylobacter* flagella. *J Bacteriol* **169**: 5066-5071.

Hennighausen, L. & Lubon, H. (1987). Interaction of protein with DNA *in vitro*. *Methods Enzymol* **152**: 721-735.

Hewlett, E.L., Sauer, K.T., Myers, G.A., Cowell, J.L. & Guerrant, R.L. (1983). Induction of a novel morphological response in Chinese hamster ovary cells by pertussis toxin. *Infect Immun* **40**: 1198-1203.

Higa, H.H. & Varki, A. (1988). Acetyl-coenzyme A: polysialic acid O-acetyltransferase from K1-positive *Escherichia coli*: the enzyme responsible for the O-acetyl plus phenotype and for O-acetyl form variation. *J Biol Chem* **263:** 8872-8878.

Higgins, C.F., Dorman, C.J., Stirling, D.A., Waddell, L., Booth, L.R., May, G. & Bremer, E. (1988). A physiological role for DNA supercoiling in the osmotic regulation of gene expression in *S. typhimurium* and *E. coli*. *Cell* **52:** 569-584.

High, N.J., Deadman, M.E. & Moxon, E.R. (1993). The role of a repetitive DNA motif (5′-CAAT-3′) in the variable expression of the *Haemophilus influenzae* lipopolysaccharide epitope αGal(1-4)ßGal. *Mol Microbiol* **9**: 1275-1282.

High, N.J., Jennings, M.P. & Moxon, E.R. (1996). Tandem repeats of the tetramer 5′-CAAT-3′ present in *lic2A* are required for phase variation but not lipopolysaccharide biosynthesis in *Haemophilus influenzae. Mol Microbiol* **20**: 165-174.

Highlander, S.K. & Garza, O. (1996). The restriction-modification system of *Pasteurella haemolytica* is a member of a new family of type I enzymes. *Gene* **178**: 89-96.

Highlander, S.K. & Hang, V.T. (1997). A putative leucine zipper activator of *Pasteurella haemolytica* leukotoxin transcription and the potential for modulation of its synthesis by slipped-strand mispairing. *Infect Immun* **65**: 3970-3975.

Hill, S.A., Morrison, S.G. & Swanson, J. (1990). The role of direct oligonucleotide repeats in gonococcal pilin gene variation. *Mol Microbiol* **4**, 1341-1352.

Hill, S.A., Belland, R.J. & Wilson, J. (1998). The *ihf* mRNA levels decline as *Neisseria gonorrhoeae* enters the stationary growth phase. *Gene* **215**: 303-310.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B-C. & Herrmann. (1996). Complete sequence analysis of the bacterium *Mycoplasma pneumoniae. Nucl Acids Res* **24**: 4420-4449.

Hitchcock, P.J., Hayes, S.F., Mayer, L.W., Shafer, W.M. & Tessier, S.L. (1985). Analysis of gonoccal H.8 antigen: surface location, inter and intra strain electorphoretic heterogeneity, and unusual two-dimensional electrophoretic characteristics. *J Exp Med* **162**: 2017-2034.

Hood D.W., Deadman, M.E., Jennings, M.P., Biscercic, M. Fleischmann, R.D., Venter, J.C. & Moxon, E.R. (1996). DNA repeats identify novel virulence genes in *Haemophilus influenzae. Proc Natl Acad Sci USA* **93**: 11121-11125.

Hopman, C.T.P., Dankert, J. & van Putten, J.P.M. (1994). Variable expression of the class 1 protein of *Neisseria meningitidis*, p. 513-517. In Conde-Glez, C.J., Morse, S., Rice, P., Sparling, F. & Calderon, E. (Ed), Pathology and immunobiology of *Neisseriaceae*. Instituto Nacional de Salud Publica Cuernavaca, Mexico.

Hsia, J.A., Moss, J., Hewlett, E.L. & Vaughan, M. (1984). ADP-ribosylation of adenylate cyclase by pertussis toxin. Effects on inhibitory agonist binding. *J Biol Chem* **259**: 1086-1090.

Hu, P.C., Cole, R.M., Huang, Y.S., Graham, J.A., Gardner, D.E., Collier, A.M. & Clyde, W.A. (1982). *Mycoplasma pneumoniae* infection: role of a surface protein in the attachment organelle. *Science* **216**: 313-315.

Idigbe, E.O., Parton, R. & Wardlaw, A.C. (1981). Rapidity of antigenic modulation of *Bordetella pertussis* in modified Hornibrook medium. *J Med Microbiol* **14**: 409-418.

Ilver, D., Kallstrom, H., Normark, S. & Jonsson, A-B. (1998). Transcellular passage of *Neisseria gonorrhoeae* involves pilus phase variation. *Infect Immun* **66**: 469-473.

Inzana, T.J., Gogolewski, R.P. & Corbeil, L.B. (1992). Phenotypic phase variation in *Haemophilus somnus* lipopolysaccharide during bovine pneumonia and after in vitro passage. *Infect Immun* **60**: 2943-2951.

Inzana, T.J., Hensley, J., McQuiston, J., Lesse, A.J., Campagnari, A.A., Boyle, S.M. & Apicella, M.A. (1997). Phase variation and conservation of lipopolysaccharide epitopes in *Haemophilus somnus*. *Infect Immun* **65**: 4675-4681.

Irons, L.I., Ashworth, L.A. & Robinson, A. (1985). Release and purification of fimbriae from *Bordetella pertussis*. *Dev Biol Stand* **61**: 153-163.

Isaacson, R.E. & Kinsel, M. (1992). Adhesion of *Salmonella typhimurium* to procine intestinal epithelial surfaces: identification and characterization of two phenotypes. *Infect Immun* **60**: 3193-3200.

Jarosik, G.P. & Hansen, E.J. (1994). Identification of a new locus involved in expression of *Haemophilus influenzae* type b lipopolysaccharide. *Infect Immun* **62**: 4861-4867.

Jarvis, G.A. (1995). Recognition and control of neisserial infection by antibody and complement. *Trends Microbiol* **3**: 198-201.

Jendrisak, J.J. & Burgess, R.R. (1975). A new method for the large-scale purification of wheat germ DNA-dependent RNA polymerase. *Biochemistry* **14**: 4639-4645.

Jennings, H.J., Bhattacharjee, A.K., Bundle, D.R., Kenny, C.P., Martin, A. & Smith, I.C. (1977). Structures of the capsular polysaccharides of *Neisseria meningitidis* as determined by ^{13}C-nuclear magnetic resonance spectroscopy. *J Infect Dis* **136**: S78-S83.

Jennings, M.P., van der Ley, P., Wilks, K.E., Maskell, D.J., Poolman, J.T. & Moxon, E.R. (1993). Cloning and molecular analysis of the *galE* gene of *Neisseria meningitidis* and its role in lipopolysaccharide biosynthesis. *Mol Microbiol* **10**: 361-369.

Jennings, M.P., Hood, D.W., Peak. I.R.A., Virji, M. & Moxon, E.R. (1995). Molecular analysis of a locus for the biosynthesis and phase-variable expression of the lacto-N-neotetraose terminal lipopolysaccharide structure in *Neisseria meningitidis*. *Mol Microbiol* **18**: 729-740.

Jennings, M.P., Virji, M., Evans, D., Foster, V., Srikhanta, Y.N., Steeghs, L., van der Ley, P. & Moxon, E.R. (1998). Identification of a novel gene involved in pilin glycosylation in *Neisseria meningitidis*. *Mol Microbiol* **29**: 975-984.

Kauffmann, F. (1954). Enterobacteriaceae, 2nd edn. Munksgaard, Copenhagen.

Kawabata, H., Myouga, F., Inagaki, Y., Murai, N. & Watanabe, H. (1998). Genetic and immunological analysis of Vls (VMP-like sequences) of *Borrelia burgdorferi*. *Microb Pathog* **24**: 155-166.

Kawula, T.H., Aho, E.L., Barritt, D.S., Klapper, D.G. & Cannon, J.G. (1988). Reversible phase variation of expression of *Neisseria meningitidis* class 5 outer membrane proteins and their relationship to gonococcal proteins II. *Infect Immun* **56**: 380-386.

Kawula, T.H. & Orndorff, P.E. (1991). Rapid site-specific DNA inversion in *Escherichia coli* mutants lacking the histone-like protein H-NS. *J Bacteriol* **173**: 4116-4123.

Kendal, W.S. & Frost, P. (1988). Pitfalls and practice of Luria-Delbrück fluctuation analysis: a review. *Cancer Res* **48**: 1060-1065.

Kim, J.J., Zhau, D., Mandrell, R.E. & Griffiss, J.M. (1992). Effects of endogenous sialylation of the lipooligosaccahride of *Neisseria gonorrhoeae* on opsonophagocytosis. *Infect Immun* **60:** 4439-4442.

Kimura, A. & Hansen, E.J. (1986). Antigenic and phenotypic variations of *Haemophilus influenzae* type b lipopolysaccharide and their relationship to virulence. *Infect Immun* **51**: 69-79.

Klein, N.J., Ison, C.A., Peakman, M., Levin, M., Hammerschmidt, S., Frosch, M. & Heyderman, R.S. (1996). The influence of capsulation and lipooligosaccharide structure on neutrophil adhesion molecule expression and endothelial injury by *Neisseria meningitidis*. *J Infect Dis* **173**: 172-179.

Klimpel, K.W., Lesley, S.A. & Clark, V.L. (1989). Identification of subunits of gonococcal RNA polymerase by immunoblot analysis: evidence for multiple sigma factors. *J Bacteriol* **171**: 3713-3718.

Knapp, S. & Mekalanos, J.J. (1988). Two trans-acting regulatory genes (*vir* and *mod*) control antigenic modulation in *Bordetella pertussis*. *J Bacteriol* **170**: 5059-5066.

Koch, A.L. (1982). Multistep kinetics: choice of models for the growth of bacteria. *J Theor Biol* **98**: 401-417.

Koch, H-G., Hwang, O. & Daldal, F. (1998). Isolation and characterisation of *Rhodobacter capsulatus* mutants affected in cytochrome cbb_3 oxidase activity. *J Bacteriol* **180**: 969-978.

Kohwi-Shigematsu, T. & Kohwi, Y. (1991). Detection of triple-helix related structures adopted by poly(dG)-poly(dC) sequences in supercoiled plasmid DNA. *Nucleic Acids Res* **19**: 4267-4271.

Koomey, M., Gotschlich, E.C., Robbins, K., Bergstrom, S. & Swanson, J. (1987). Effects of *recA* mutations on pilus antigenic variation and phase transitions. *Genetics* **117**: 391-398.

Krantz, I., Alestig, K., Trollfors, B. & Zackrisson, G. (1986). The carrier state in pertussis. *Scand J Infect Dis* 1986 **18**: 121-123.

Jennings, M.P., Srithanta, Y.N., Moxon, E.R., Kramer, M., Poolman, J.T., Kuipers, B. & van der Ley, P. (1999). The genetic basis of the phase variation repertoire of lipopolysaccharide immunotypes in *Neisseria meningitidis*. *Microbiology* **145:** 3013-3021.

Jerse, A.E., Cohen, M.S., Drown, P.M., Whitaker, L.G., Isbey, S.F., Seifert, H.S. & Cannon,J.G. (1994). Multiple gonococcal opacity proteins are expressed during experimental urethral infection in the male. *J Exp Med* **179:** 911-920.

Johnson, E.M. & Baron, L.S. (1969). Genetic transfer of the Vi antigen from *Salmonella typhosa* to *Escherichia coli*. *J Bacteriol* **99**: 355-359.

Johnson, J.R. (1991). Virulence factors in *Escherichia coli* urinary tract infection. *Clin Microbiol Rev* **4**: 80-125.

Jones, S.A., Marchitto, K.S., Miller, J.N. & Norgard, M.V. (1984). Monoclonal antibody with hemaggluination, immobilisation, and neutralization activities defines an immunodominant, 47,000 mol weight, surface exposed immunogen of *Treponema pallidum*. *J Exp Med* **160**: 1404-1420.

Jones, D.M., Borrow, R., Fox, A.J., Gray, S., Cartwright, K.A. & Poolman, J.T. (1992). The lipooligosaccharide immunotype as a virulence determinant in *Neisseria meningitidis*. *Microb Pathog* **13**: 219-224.

Jones, M.E., Thomas, S.M. & Rogers, A. (1994). Luria-Delbruck fluctuation experiments: design and analysis. *Genetics* **136**: 1209-1216.

Jonsson, A.B., Nyberg, G. & Normark, S. (1991) Phase variation by gonococcal pili by frameshift mutation in *pilC*, a novel gene for pilus assembly. *EMBO J* **10**, 477-488.

Judd, R.S. & Shafer, W.M. (1989). Topographical alterations in proteins I of *Neisseria gonorrhoeae* correlated with lipooligosaccharide variation. *Mol Microbiol* **3**: 637-642.

Kamp, D., Kahmann, R., Zipser, D., Broker, T.R. & Chow, L.T. (1978). Inversion of the G DNA segment of phage Mu controls phage infectivity. *Nature* **271**:577-580.

Kasper, D.L., Winkelhake, J.L., Zollinger, W.D., Brandt, B.L. & Artenstein, M.S. (1973). Immunological similarity between polysaccharide antigens of *Escherichia coli* O7:K1 (L): NM and group B *Neisseria meningitidis*. *J Immunol* **110**: 262-268.

Kasper, D.L., Baker, C.J., Galdes, B., Katzenellenbogen, E. & Jennings, H.J. (1983). Immunologica analysis and immunogenicity of the type II group B streptococccal capsular polysaccharide. *J Clin Invest* **72**: 260-269.

Katada, T. & Ui, M. (1982). Direct modification of the membrane adenylate cyclase system by islet-activating protein due to ADP-ribosylation of a membrane protein. *Proc Natl Acad Sci USA* **79**: 3129-3133.

Krause, D.C., Leith, D.K., Wilson, R.M. & Baseman, J.B. (1982). Identificaion of *Mycoplasma pneumoniae* proteins associated with hemadsorption and virulence. *Infect Immun* **35**: 809-817.

Krause, D.C., Leith, D.K. & Baseman, J.B. (1983). Reaquisition of specific proteins confers virulence in *Mycoplasma pneumoniae*. *Infect Immun* **39**: 830-836.

Kulasekara, H.D. & Blomfield, I.C. (1999). The molecular basis for the specificity of *fimE* in the phase variation of type 1 fimbriae of *Escherichia coli* K-12. *Mol Microbiol* **31**: 1171-1181.

Kupsch, E-M., Knepper, B., Kuroki, T., Heuer, I. & Meyer, T.F. (1993). Variable opacity (Opa) outer membrane proteins account for the cell tropisms displayed by *Neisseria gonorrhoeae* for human leukocytes and epithelial cells. *EMBO J* **12**: 641-650.

Kwan, L.Y. & Isaacson, R.E. (1998). Identification and characterisation of a phase-variable nonfimbrial *Salmonella typhimurium* gene that alters O-antigen production. *Infect Immun* **66**: 5725-5730.

Lacey, B.W. (1960). Antigenic modulation of *Bordetella pertussis*. *J. Hyg* **58**: 57-93.

Lambden, P.R., Heckles, J.E., James, L.T. & Watt, P.J. (1979). Variations in surface protein composition associated with virulence properties in opacity types of *Neisseria gonorrhoeae*. *J Gen Microbiol* **114**: 305-312.

Lambden, P.R., Robertson, J.N. & Watt, P.J. (1980). Biological properties of two distinct pilus types produced by isogenic variants of *Neisseria gonorrhoeae* P9. *J Bacteriol* **141**: 393-396.

Langermann, A. & Wright, A. (1990). Molecular analysis of the *Haemophilus influenzae* type b pilin gene. *Mol Microbiol* **4**: 221-230.

Lea, D.E. & Coulson, C.A. (1949). The distribution of the numbers of mutants in bacterial populations. J *Genetics* **49**: 264-285.

LeClerk, J.E., Li, B., Payne, W.L. & Cebula, T.A. (1996). High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science* 274, 1208-1211.

Lederberg, J. & Edwards, P.R. (1953). Serotypic recombination in *Salmonella*. *J Immunol* **71**: 232-240.

Lederberg, J. & Iino, T. (1956). Phase variation in *Salmonella*. *Genetics* **41**: 743-757.

Leininger, E., Roberts, M., Kenimer, J.G., Charles, I.G., Fairweather, N., Novotny, P. & Brennan, M.J. (1991). Pertactin, an Arg-Gly-Asp-containing *Bordetella pertussis* surface protein that promotes adherence of mammalian cells. *Proc Natl Acad Sci USA* **88**: 345-349.

Lenich, A.G. & Glasgow, A.C. (1994). Amino acid sequence homology between Piv, an essential protein in site-specific DNA inversion in *Moraxella lacunata*, and transposases of an unusual family of insertion elements. *J Bacteriol* **176**: 4160-4164.

Lerbs, S., Brautigam, E. & Parthier, B. (1985). Polypeptides of DNA-dependent RNA polymerase of spinach chloroplasts: characterisation by antibody-linked polymerase assay and determination of sites of synthesis. *EMBO J* **4**: 1661-1666.

Levinson, G. & Gutman, G.A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 4: 203-221.

Lewis, L.A., Gipson, M., Hartman, K., Ownbey, T., Vaughn, J. & Dyer, D.W. (1999). Phase variation of HpuAB and HmbR, two distinct haemoglobin receptors of *Neisseria meningitidis* DNM2. *Mol Microbiol* **32**: 977-989.

Li, I-c. & Chu, E.H.Y. (1987). Evaluation of methods for the estimation of mutation rates in cultured mammalian cell populations. *Mutation Res* **190**: 281-287.

Lim, J.K., Gunther, N.W., Zhao, H., Johnson, D.E., Keay, S.K. & Mobley, H.L.T. (1998). *In vivo* phase variation of *Escherichia coli* type 1 fimbrial genes in women with urinary tract infection. *Infect Immun* **66**: 3303-3310.

LiPuma, J. & Gilsdorf, J.R. (1988). Structural and serological relatedness of *Haemophilus influenzae* type b pili. *Infect Immun* **56**: 1051-1056.

Locht, C. & Keith, J.M. (1986). Pertussis toxin gene: nucleotide sequence and genetic organization. *Science* **232**: 1258-1264.

Locht, C., Bertin, P., Menozzi, F.D. & Renauld, G. (1993). The filamentous haemagglutinin, a multifaceted adhesin produced by virulent *Bordetella* spp. *Mol Microbiol* 9: 653-660.

Lowe, P.A., Hager, D.A. & Burgess, R.R. (1979). Purification and properties of the σ subunit of *Escherichia coli* DNA-dependent RNA polymerase. *Biochemistry* **18**: 1344-1352.

Luria, S.E. & Delbrűk M. (1943). Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**: 491-511.

Lynch, E.C., Blake, M.S., Gotschlich, E.C. & Mauro, A. (1984). Studies of porins – spontaneously transferred from whole cells and reconstituted from purified proteins of *Neisseria gonorrhoeae* and *Neisseria meningitidis*. *Biophys J* **45**: 104-107.

Lysnyansky, I., Rosengarten, R. & Yogev, D. (1996). Phenotypic switching of variable surface lipoproteins in *Mycoplasma bovis* involves high-frequency chromosomal rearrangements. *J Bacteriol* **178**: 5395-5401.

Mackinnon, F.G., Borrow, R., Gorringe, A.R., Fox, A.J., Jons, D.M. & Robinson, A. (1993). Demonstration of lipopolysaccharide immunotype and capsule as virulence factors for *Neisseria meningitidis* using an infant mouse intranasal infection model. *Microb Pathog* **15**: 359-366.

MacNab, R.M. (1996). Flagella and motility. In. Neidhardt, F.C. Ed. *Escherichia coli* and *Salmonella*. 2nd Ed. Chapter 10, 123-145.

Magnuson, H.J., Thomas, E.W., Olansky, S., Kaplan, B.I., DeMello, L. & Cutler, J.C. (1956). Inoculation syphilis in human volunteers. *Medicine (Baltimore)* **35**: 33-82.

Makino, S-I., van Putten, J.P.M. & Meyer, T.F. (1991). Phase variation of the opacity outer membrane protein controls invasion by *Neisseria gonorrhoeae* into human epithelial cells. *EMBO J* **10**: 1307-1315.

Manning, D.R., Fraser, B.A., Kahn, R.A. & Gilman, G.A. (1984). ADP-ribosylation of transducin by islet-activation protein. Identification of asparagine as the site of ADP-ribosylation. *J Biol Chem* **259**: 749-756.

Manning, P.A., Kaufmann, A., Roll, U., Pohlner, J., Meyer, T.F. & Haas, R. (1991). L-pilin variants of *Neisseria gonorrhoeae* MS11. *Mol Microbiol* **5**: 917-926.

Manning, D.S., Reschke, D.K. & Judd, R.C. (1998). Omp85 proteins of *Neisseria gonorrhoeae* and *Neisseia meningitidis* are similar to *Haemophilus influenzae* D-15-Ag and *Pasteurella multocida* Oma87. *Microb Pathog* 25: 11-21.

Marceau, M., Beretti, J-L. & Nassif, X. (1995). High adhesiveness of encapsulated *Neisseria meningitidis* to epithelial cells is associated with the formation of bundles of pili. *Mol Microbiol* **17**: 855-863.

Marceau, M., Forest, K., Beretti, J-L., Tainer, J. & Nassif, X. (1998). Consequence of the loss of O-linked glycosylation of meningococcal type IV pilin for piliation and pilus mediated adhesion. *Mol Microbiol* **27**: 705-715.

Marchitto, K.S., Jones, S.A., Schell, R.F., Holmans, P.L. & Norgard, M.V. (1984). Monoclonal antibody analysis of specific antigenic similarities among pathogenic *Treponema pallidum* subspecies. *Infect Immun* **45**: 660-666.

Marchitto, K.S., Selland-Grossling, C.K. & Norgard, M.V. (1986). Molecular specificities of monoclonal antibodies directed against virulent *Treponema pallidum*. *Infect Immun* **51**: 168-176.

Marrs, C.F., Schoolnik, G., Koomey, J.M., Hardy, J., Rothbard, J. & Falkow, S. (1985). Cloning and sequencing of a *Moraxella bovis* pilin gene. *J Bacteriol* **163**: 132-139.

Marrs, C.F., Ruehl, W.W., Schoolnik, G.K. & Falkow, S. (1988). Pilin gene phase variation of *Moraxella bovis* is caused by an inversion of the pilin genes. *J Bacteriol* **170**: 3032-3039.

Marrs, C.F., Rozsa, F.W., Hackel, M., Stevens, S.P. & Glasgow, A.C. (1990). Identification, cloning, and sequencing of piv, a new gene involved in inverting the pilin genes of *Moraxella lacunata*. *J Bacteriol* **172**: 4370-4377.

Maskell, D.J., Szabo, M.J., Butler, P.D., Williams, A.E. & Moxon, E.R. (1991). Molecular analysis of a complex locus from *Haemophilus influenzae* involved in phase-variable lipopolysaccharide biosynthesis. *Mol Microbiol* **5**: 1013-1022.

Maskell, D.J., Szabo, M.J. & High, N.J. (1993). PCR amplification of DNA sequences from nitrocellulose-bound, immunostained bacterial colonies. *Nucl Acids Res* **21**: 171-172.

Mason, E.O., Jr., Kaplan, S.L., Weidermann, B.L., Norrod, E.P. & Stenback, W.A. (1985). Frequency and properties of naturally occuring adherent piliated strains of *Haemophilus influenzae* type b. *Infect Immun* **49**: 98-103.

McClure, W.R. (1985). Mechanism and control of transcription initiation in prokaryotes. *Annu Rev Biochem* **54**: 171-204.

McGuinness, B.T., Clarke, I.N., Lambden, P.R., Barlow, A.K., Poolman, J.T., Jones D.M., & Heckels J.E. (1991). Point mutation in meningocccal porA gene associated with increased endemic disease. *Lancet* **337**: 514-517.

McNeil, G., Virji, M. & Moxon, E.R. (1994). Interactions of *Neisseria meningitidis* with human monocytes. *Microb Pathog* **16**: 153-163.

McNeil, G. & Virji, M. (1997). Phenotypic variants of meningococci and their potential in phagocytic interactions: the influence of opacity proteins, pili, PilC and surface sialic acids. *Microl Pathog* **22**: 295-304.

Meier, J.T., Simon, M.I. & Barbour, A.G. (1985). Antigenic variation is associated with DNA rearrangements in a relapsing fever *Borrelia*. *Cell* **41**: 403-409.

Meleney, H.E. (1928). Relapse phenomena of *Spironema recurrentis*. *J Exp Med* **48**: 65-82.

Merker, P., Tommassen, J., Kusecek, B., Virji, M., Sesardic, D. & Achtman, M. (1997). Two-dimensional structure of the Opc invasin from *Neisseria meningitidis*. *Mol Microbiol* **23**: 281-293.

Mertsola, J., Cope, L.D., Saez-Lorenz, X., Ramilo, O., Kennedy, W., McCraken, G.H., Jr & Hansen, E.J. (1991). *In vivo* and *in vitro* expression of *Haemophilus influenzae* type b lipooligosaccharide epitopes. *J Infect Dis* **164**: 555-563.

Messer, W. & Weigel, C. (1996). Initiation of chromosomal replication. pages: 1579-1601. In: *Escherichia coli* and *Salmonella* (second edition). Ed: Neidhardt, F.C. ASM Press.

Meyer, T.F., Billyard, E., Haas, R., Storzbach, S. & So, M. (1984). Pilus genes of *Neisseria gonorrhoeae*: chromosomal organisation and DNA sequence. *Proc Natl Acad Sci USA* **81**: 6110-6114.

Miller, V.L., Taylor, R.K. & Mekalanos, J.J. (1987). Cholera toxin transcriptional activator ToxR is a transmembrane DNA binding protein. *Cell* **48**: 271-279.

Mittler, J.E. & Lenski, R.E. (1992). Experimental evidence for an alternative to directed mutation in the *bgl* operon. *Nature* **356**: 446-448.

Morrison-Plummer, J., Leith, D.K. & Baseman, J.B. (1986). Biological effects of anti-lipid and anti-protein monoclonal antibodies of *Mycoplasma pneumoniae. Infect Immun* **53**: 398-403.

Moxon, E.R., Rainey, P.B., Nowak, M.A. & Lenski, R.E. (1994). Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol* **4**, 24-33.

Moxon, E.R., Gewurz, B.E., Richards, J.C., Inzana, T., Jennings, M.P. & Hood, D.W. (1996). Phenotypic switching of *Haemophilus influenzae. Mol Microbiol* **19**: 1149-1150.

Moxon, E.R. & Thaler, D.S. (1997). The tinkerer's evolving tool-box. *Nature* **387**, 659-662.

Mulligan, M.E., Brosius, J. & McClure, W.R. (1985). Characterization *in vitro* of the effect of spacer length on the activity of Escherichia coli RNA polymerase at the TAC promoter. *J Biol Chem* **260**: 3529-3538.

Munkley, A., Tinsley, .R., Virji, M. & Heckels, J.E. (1991). Blocking of bactericidal killing of *Neisseria meningitidis* by antibodies directed against class 4 outer membrane protein. *Microb Pathog* **11**: 447-452.

Munoz, J.J., Arai, H. & Cole, R.L. (1981a). Mouse protecting and histamine-sensitizing activities of pertussigen and fimbrial hemagglutinin from *Bordetella pertussis*. *Infect Immun* **32**: 243-250.

Munoz, J.J., Arai, H., Bergman, R.K. & Sadowski, P.L. (1981b). Biological activities of crystalline pertussigen from *Bordetella pertussis*. *Infect Immun* **33**: 820-826.

Murphy, G.L., Connell, T.D., Barritt, D.S., Koomey, M. & Cannon J.G. (1989). Phase variation of gonococcal protein II: regulation of gene expression by slipped-strand mispairing of a repetitive DNA sequence. *Cell* **56**: 539-547.

Nassif, X., Lowy, J., Stenberg, P., O'Gaora, P., Ganji, A. & So, M. (1993). Antigenic variation of pilin reulates adhesion of *Neisseria meningitidis* to human epithelial cells. *Mol Microbiol* **8**: 719-725.

Nassif, X., Beretti, J-L., Lowy, J., Stenberg, P., O'Gaora, P., Pfeifer, J., Normark, S. & So, M. (1994). Roles of pilin and PilC in adhesion of *Neisseria meningitidis* to human epithelial and endothelial cells. *Proc Natl Acad Sci USA* **91**: 3769-3773.

Naumann, M., Rudel, T. & Meyer, T.F. (1999). Host cell interactions and signalling with *Neisseria gonorrhoeae. Curr Opin Microbiol* **2**: 62-70.

Newcombe, H.B. (1948). Delayed phenotypic expression of spontaneous mutations in *Escherichia coli*. *Genetics* **33**: 447-476.

Nicholson, A. & Lepow, I.H. (1979). Host defecnse against *Neisseria meningitidis* requires a complement-dependent bactericidal activity. *Science* **205**: 298-299.

Nicosia, A. & Rappuoli, R. (1987). Promoter of the pertussis toxin operon and production of pertussis toxin. *J Bacteriol* **169**: 2843-2846.

Nicosia, A., Perugini, M., Franzini, C., Casagli, M.C., Borri, M.G., Antoni, G., Almoni, M., Neri, P., Ratti, Giulio. & Rappuoli, R. (1986). Cloning and sequencing of the pertussis toxin genes: operon structure and gene duplication. *Proc Natl Acad Sci USA* **83**: 4631-4635.

Nikaido, H. (1996). Outer membrane. pages: 29-47. In: *Escherichia coli* and *Salmonella* (second edition). Ed: Neidhardt, F.C. ASM Press.

Norgard, M.V., Marchitto, K.S. & Cox, D.L. (1986). Phehotypic expression of the major 47 kDa surface immunogen of *Treponema pallidum* in virulent, tissue-cultured treponemes. *J Gen Microbiol* **132**: 1775-1778.

Norlander, L., Davies, J., Norqvist, A. & Normark, S. (1979). Genetic basis for colonial variation in *Neisseria gonorrhoeae*. *J Bacteriol* **138**: 762-769.

Norris, S.J. (1982). In vitro cultivation of *Treponema pallidum*: independent confirmation. *Infect Immun* **36**, 437-439.

Norris, S.J. and the Treponema pallidum polypeptide research group. (1993). Polypeptides of *Treponema pallidum*: progress toward understanding their structural, functional, and immunologic roles. *Microbiol Rev* **57**, 750-779.

Nou, X., Skinner, B., Braaten, B., Blyn, L., Hirsh, D. & Low, D. (1993). Regulation of pyelonephritis-associated pili phase variation in *Escherichia coli*: binding of the PapI and Lrp regulatory proteins is controlled by DNA methylation. *Mol Microbiol* 7: 545-553.

Nou, X., Braaten, B., Kaltenbach, L. & Low, D.A. (1995). Differential binding of Lrp to two sets of *pap* DNA binding sites mediated by PapI regulates Pap phase variation in *Escherichia coli*. *EMBO J* **14**: 5785-5797.

Ofek, I. & Beachey, E.H. (1978). Mannose binding and epithelial cell adherence of *Escherichia coli*. *Infect Immun* **22**: 247-254.

Ogle, K.F., Lee, K.K. & Krause, D.C. (1991). Cloning and analysis of the gene encoding the cytadherence phase-variable protein HMW3 from *Mycoplasma pneumoniae*. *Gene* **97**: 69-75.

Olyhoek, A.J. Sarkari, J., Bopp, M., Morelli, G. & Achtman, M. (1991). Cloning and expression in *Escherichia coli* of *opc*, the gene for an unusual class 5 outer membrane protein from *Neisseria meningitidis*. *Microb Pathog* **11**: 249-257.

Orskov, F., Orskov, I., Sutton, A., Schneerson, R., Lin, W., Egan, W., Hoff, G.E. & Robbins, J.B. (1979). Form variation in *Escherichia coli* K1: determined by O-acetylation of the capsular polysaccharide. *J Exp Med* **149**: 669-685.

Ou, J.T., Baron, L.S., Rubin, F.A. & Kopecko, D.J. (1988). Specific insertion and deletion of insertion sequence 1-like DNA element causes the reversible expression of the virulence capsular antigen Vi of *Citrobacter freundii* in *Escherichia coli*. *Proc Natl Acad Sci USA* **85**: 4402-4405.

Parge, H.E., Forest, K.T., Hickey, M.J., Christensen, D.A., Getzoff, E.D. & Tainer, J.A. (1995). Structure of the fibre-forming protein pilin at 2.6 Å resolution. *Nature* **378**: 32-38.

Parsons, N.J., Andrade, J.R.C., Patel, P.V., Cole, J.A. & Smith, H. (1989). Sialylation of lipopolysaccharide and loss of absorption of bactericidal antibody during conversion of gonococci to serum resistance by cytidine 5′-monophospho-N-acetyl neuraminic acid. *Microb Pathog* **7**: 63-72.

Parsons, N.J., Cole, J.A. & Smith, H. (1990). Resistance to human serum of gonococci in urethral exudates is reduced by neuraminidase. *Proc R Soc Lond B* **241**: 3-5.

Paruchuri, D.K. & Harshey, R.M. (1987). Flagellar variation in *Serratia marcescens* is associated with color variation. *J Bacteriol* **169**: 61-65.

Patrick, C.C., Patrick, G.S., Kaplan, S.L., Barrish, J. & Mason, E.O., Jr. (1989). Adherence kinetics of *Haemophilus influenzae* type b to eukaryotic cells. *Pediatr Res* **26**: 500-503.

Peak, I.R.A., Jennings, M.P., Hood, D.W., Bisercic, M. & Moxon, E.R. (1996). Tetrameric repeat units associated with virulence factor phase variation in Haemophilus also occur in *Neisseria* spp. and *Moraxella catarrhalis*. *FEMS Microbiol Lett* **137**: 109-114.

Peak, I.R.A., Jennings, M.P., Hood, D.W. & Moxon, E.R. (1999). Tetranucleotide repeats identify novel virulence determinant homologues in *Neisseria meningitidis*. *Microb Pathog*: **26**: 13-23.

Pennisi, E. (1998) Genome reveals wiles and weak points of syphilis. *Nature* **281**, 324-325.

Peppler, M.S. (1984). Two physically and serologically distinct lipopolysaccharide profiles of *Bordetella pertussis* and their phenotypic variants. *Infect Immun* **43**: 224-232.

Peppler, M.S. & Schrumph, M.E. (1984). Phenotypic variation and modulation in *Bordetella bronchiseptica*. *Infect Immun* **44**: 681-687.

Perry, A.C.F., Nicolson, I.J. & Saunders, J.R. (1987). Structural analysis of the *pilE* region of *Neisseria gonorrhoeae* P9. *Gene* **60**: 85-92.

Peterson, K.M., Baseman, J.B. & Alderete, J.F. (1986). Isolation of a *Treponema pallidum* gene encoding immunodominant outer envelope protein P6, which reacts with sera from patients at different stages of syphilis. *J Exp Med* **164**: 1160-1170.

Pettersson, A., van der Ley, P., Poolman, J.T. & Tommassen, J. (1993). Molecular characterisation of the 98-kilodalton iron-regulated outer membrane protein of *Neisseria meningitidis*. *Infect Immun* **61**: 4724-4733.

Pettersson, A., Maas, A., van Wassenaar, D., van der Ley, P. & Tommassen, J. (1995). Molecular characterization of FrpB, the 70-kildalton iron-regulated outer membrane protein of *Neisseria meningitidis*. *Infect Immun* **63**: 4181-4184.

Pettit, R.K., Szuba, J.C. & Judd, R.C. (1990). Characterisation of fourteen strains of *Neisseria gonorrhoeae*: stuctural analysis and serum reactivities. *Mol Microbiol* 4: 1293-1301.

Pichichero, M.E., Anderson, P., Loeb, M. & Smith, D.H. (1982). Do pili play a role in pathogenicity of *Haemophilus influenzae* type b? *Lancet* **2**: 960-962.

Pincus, S.H., Cole, R.L., Wessels, M.R., Corwin, M.D., Kamanga-Sollo, E., Hayes, S.F., Cieplak, W. Jr. & Swanson, J. (1992). Group B streptococcal opacity variants. *J Bacteriol* **174**: 3739-3749.

Pitman, M. (1931). Variation and type specificity in the bacterial species *Haemophilus influenzae*. *J Exp Med* **53**: 471-492.

Plasterk, R.H.A., Simon, M.I. & Barbour, A.G. (1985). Transposition of structural genes to an expression sequence on a linear plasmid causes antigenic variation in the bacterium *Borrelia hermsii*. *Nature* **318**: 257-263.

Poolman, J.T., de Marie, S. & Zanen, H.C. (1980). Variability of low-molecular-weight, heat-modifiable outer membrane proteins of *Neisseria meningitidis*. *Infect Immun* **30**: 642-648.

Poolman, J.T., Timmermans, H.A.M., Hopman, C.T.P., Teerlink, T., van Vught, P.A.M., Witvliet, M.H. & Beuvery, E.C. (1988). in Poolman, J.T., Znen, H.C., Meyer, T.F., Heckels, J.E., Makela, P.R.H., Smith, H. & Bauvery, E.C. (eds). Gonococci and Meningococci. Klewer Acdemic, Dortrecht: 159-166.

Porat, N., Apicella, M.A. & Blake, M.S. (1995). A lipopolysaccharide-binding site on HepG2 cells similar to the gonococcal opacity-associated surface protein Opa. *Infect Immun* **63**: 2164-2172.

Preisig, O., Anthamatten, D. & Hennecke, H. (1993). Genes for a microaeophilically induced oxidase complex in *Bradyrhizobium japonicum* are essential for a nitogen-fixing endosymbiosis. *Proc Natl Acad Sci USA*: **90**: 3309-3313.

Preisig, O., Zufferey, R., Thony-Meyer, L., Appleby, C.A. & Hennecke, H. (1996). A high-affinity cbb_3-type cytochrome oxidase terminates the symbiosis-specific respiratory chain of *Bradyrhizobium japonicum*. *J Bacteriol* **178**: 1532-1538.

Preston, N.W., Timewell, R.M. & Carter, E.J. (1980). Experimental pertussis infection in the rabbit: similarities with infection in primates. *J Infect* **2**: 227-235.

Pron, B., Taha, M-K., Rambaud, C., Fournet, J-C., Pattey, N., Monnet J-P., Musilek, M., Beretti, J-L. & Nassif, X. (1997). Interaction of *Neisseria meningitidis* with the components of the blood-brain barrier correlates with an increased expression of PilC. *J Infect Dis* **176**: 1285-1292.

Radolf, J.D., Norgard, M.V. & Schulz, W.W. (1989). Outer membrane ultrastructure explains the limited antigenicity of virulent *Treponema pallidum*. *Proc Natl Acad Sci USA* **86**, 2051-2055.

Rahman, M., Kallstrom, H., Normark, S. & Jonsson, A. (1997). PilC of pathogenic *Neisseria* is associated with the bacterial cell surface. *Mol Microbiol* **25**: 11-25.

Rayner, C.F.J., Dewar, A., Moxon, E.R., Virji, M. & Wilson, R. (1995). The effect of variations in the expression of pili on the interaction of *Neisseria meningitidis* with human nasopharyngeal epithelium. *J Infect Dis* **171**: 113-121.

Read, R.C., Wilson, R., Rutman, A., Lund, V., Todd, C.H., Brain, A.P.R., Jeffrey, P.K. & Cole, P.J. (1991). Interaction of nontypable *Haemophilus influenzae* with human respiratory mucosa *in vitro*. *J Infect Dis* **163**: 549-558.

Read, R.C., Zimmerli, S., Broaddus, V.C., Sanan, D.A., Stephens, D.S. & Ernst, J.D. (1996a). The (α2→8)-linked polysialic acid capsule of group B *Neisseria meningitidis* modifies multiple steps during interaction with human macrophages. *Infect Immun* **64**: 3210-3217.

Read, T.D., Dowdell, M., Satola, S.W. & Farley, M.M. (1996b). Duplication of pilus gene complexes of *Haemophilus influenzae* biogroup aegyptius. *J Bacteriol* **178**: 6564-6570.

Relman, D.A., Domenghini, M., Tuomanen, E., Rappuoli, R. & Falkow, S. (1989). Filamentous hemagglutinin of *Bordetella pertussis*: necleotide sequence and crucial role in adherence. *Proc Natl Acad Sci USA* **86**: 2637-2641.

Rest, R.F. & Frangipane, J.V. (1992). Growth of *Neisseria gonorrhoeae* in CMP-N-acetylneuraminic acid inhibits nonopsonic (opacity-associated outer membrane protein-mediated) interactions with human neutrophils. *Infect Immun* **60**: 989-997.

Richardson, A.R. & Stojiljkovic, I. (1999). HmbR, a hemoglobin-binding outer membrane protein of *Neisseria meningitidis*, undergoes phase variation. *J Bacteriol* **181**: 2067-2074

Risberg, A., Schweda, E.K.H. & Jansson, P.-E. (1997). Stuctural studies of the cell-envelop oligosaccharide from lipopolysaccharide of *Haemophilus influenzae* strain RM 118-28. *Eur J Biochem* **243**: 701-707.

Robertson, P.W., Goldberg, H., Jarvie, B.H., Smith, D.D. &Whybin, L.R. (1987). *Bordetella pertussis* infection: a cause of persistent cough in adults. *Med J Aust* **147**: 5222-5225.

Robinson, A. & Irons, L.I. (1983). Synergistic effect of *Bordetella pertussis* lymphocytosis-promoting factor on protective activities of isolated *Bordetella* antigens in mice. *Infect Immun* **40**: 523-528.

Robinson, E.N., Jr, McGee, Z.A., Buchanan, T.M., Blake, M.S. & Hitchcock, P.J. (1987). Probing the surface of *Neisseria gonorrhoeae*: simultaneous localisation of protein I and H.8 antigens. *Infect Immun* **55**: 1190-1197.

Roche, R.J., High, N.J. & Moxon, E.R. (1994). Phase variation of *Haemophilus influenzae* lipopolysaccharide: characterization of lipopolysaccharide from individual colonies. *FEMS Microbiol Lett* **120**: 279-284.

Roche, R.J. & Moxon, E.R. (1995). Phenotypic variation in *Haemophilus influenzae*: The interrelationship of colony opacity, capsule and lipopolysaccharide. *Microb Pathog* **18**: 129-140.

Rosengarten, R. & Wise, K.S. (1990). Phenotypic switching in Mycoplasmas: Phase variation of diverse surface lipoproteins. *Science* **247**: 315-318.

Rosengarten, R. & Wise, K.S. (1991). The Vlp system of *Mycoplasma hyorhinis*: combinatorial expression of distinct size variant lipoproteins generating high-frequency surface antigenic variation. *J Bacteriol* **173**: 4782-4793.

Rosengarten, R., Behrens, A., Stetefeld, A., Heller, M., Ahrens, M., Sachse, K., Yogev, D. & Kirchhoff, H. (1994). Antigenic heterogeneity among isolates of *Mycoplasma bovis* is generated by high-frequency variation of diverse membrane surface proteins. *Infect Immun* **62**: 5066-5074.

Rosenqvist, E., Hoiby, E.A., Wedege, e., Kusecek, B. & Achtman, M. (1993). The 5C protein of *Neisseria meningitidis* is highly immunogenic in humans and induces bactericidal antibodies. *J Infect Dis* **167**: 1065-1073.

Rosenqvist, E., Høiby, E.A., Wedege, E., Bryn, K., Kolberg, J., Klem, A., Rønnild, E., Bjune, G. & Nøkleby, H. (1995). Human antibody responses to meningococcal outer membrane antigens after three doses of the Norwegian group B meningococcal vaccine. *Infect Immun* **63**: 4642-4652.

Rozsa, F.W. & Marrs, C.F. (1991). Interesting sequence differences between the pilin gene inversion regions of *Moraxella lacunata* ATCC 17956 and *Moraxella bovis* Epp63. *J Bacteriol* **173**: 4000-4006.

Rudel, T., van Putten, J.P.M., Gibbs, C.P., Haas, R. & Meyer, T.F. (1992). Interaction of two variable proteins (PilE and PilC) required for pilus mediated adherence of *Neisseria gonorrhoeae* to human epithelial cells. *Mol Microbiol* **6**: 3439-3450.

Rudel, T., Scheuerpflug, I. & Meyer, T.F. (1995a). *Neisseria* PilC protein identified as type-4 pilus tip-located adhesin. *Nature* **373**: 357-362.

Rudel, T., Boxberger, H-J. & Meyer, T.F. (1995b). Pilus biogenesis and epithelial adherence of *Neisseria gonorrhoeae pilC* double knock-out mutants. *Mol Microbiol* **17**: 1057-1071.

Rudel, T., Facius, D., Barten, R., Scheuerpflug, I., Nonnenmacher, E. & Meyer, T.F. (1995c). Role of pili and phase variable PilC protein in natural competence for transformation of *Neisseria gonorrhoea*. *Proc Natl Acad Sci USA* **92**: 7986-7990.

Ryan, K.A. & Lo, R.Y.C. (1999). Characterization of a CACAG pentanucleotide repeat in *Pasteurella haemolytica* and its possible role in modulation of a novel type III restriction-modification system. *Nucleic Acids Res* **27**: 1505-1511.

Salit , I.E. & Gotschlich, E.C. (1977). Type I *Escherichia coli* pili: characterization of binding to monkey kidney cells. *J Exp Med* **146**: 1182-1194.

Sambrook, J., Fritsch, E.F. & Maniatis, T. (1989). Molecular Cloning – A laboratory manual. Cold Spring Harbour Laboratory Press.

Sanger F, Nicklen S and Coulson A R (1977). DNA sequencing with chain-terminating inhibitors. *Proc Nat Acad Sci USA* **74**: 5463-5467.

Sarkari, J.F., Olyhoek, T., Bopp, M., *et al.* (1991). Class 5 proteins and *opa* genes in clones IV-1 and III-1 of *Neisseria meningitidis* serogroup A. In: Actman, M., Kohl, P., Marchal, C., *et al.*, Eds. Neisseria 1990. Walter de Gruyter & Co. Berlin, 1991: 539-544.

Sarkari, J., Pandit, N., Moxon, E.R. & Achtman, M. (1994). Variable expression of the Opc outer membrane protein in *Neisseria meningitidis* is caused by size variation of a promoter containing poly-cytidine. *Mol Microbiol* **13**: 207-217.

Sarker, S., Ma, M.T. & Sandri, H. (1992). On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. *Genetica* **85**: 173-179.

Sato, H. & Sato, Y. (1984). *Bordetella pertussis* infection in mice: correlation of specific antibodies against two antigens, pertussis toxin, and filamentous hemagglutinin with mouse protectivity in an intracerebral or aerosol challenge system. *Infect Immun* **46**: 415-421.

Saukonen, K., Abdillahi, H., Poolman, J.T. & Leinonen, M. (1987). Protective efficacy of monoclonal antibodies to class 1 and class 3 outer membrane proteins of *Neisseria meningitidis* B15:P1.16 in infant rat infection model: new prospects for vaccine development. *Microb Pathog* **3**: 261-267.

Saukonen, K., Abdillahi, H., Poolman, J.T. & Leinonen, M. (1989). Comparative evaluation of potential components for group B meningococcal vaccine by passive protection in the infant rat and *in vitro* bactericidal assay. *Vaccine* **7**: 325-328.

Saunders, N.J., Peden, J.F., Hood, D.W. & Moxon, E.R. (1998). Simple sequence repeats in the *Helicobacter pylori* genome. *Mol Microbiol* **27**: 1091-1098

Saunders, N.J. & Moxon, E.R. (1998). Implications of sequencing bacterial genomes for pathogenesis and vaccine development. *Curr Opinion Biotech* **9**, 618-623.

Saunders, N.J., Hood, D.W. & Moxon, E.R. (1999a). Bacteria play pass the gene. *Curr Biol* **9**: R180-R183.

Saunders, N.J., Peden, J.F. & Moxon, E.R. (1999b). The absence of an uptake sequence in *Helicobacter pylori*. *Microbiology* **145**: 3523-3528.

Savery, N.J., Rhodius, V.A., Wing, H.J. & Busby, S.J.W. (1995). Transcription activation at *Escherichia coli* promoters dependent on the cyclic AMP receptor protein: effects of binding sequences for the RNA polymerase α-subunit. *Biochem J* **309**: 77-83.

Schneider, G.J., Tumer, N.E., Richaud, C., Borbely, G. & Haselkorn, R. (1987). Purification and characterisation of RNA polymerase from the cyanobacterium *Anabaena* 7120. *J Biol Chem* **262**: 14633-14639.

Schneider, H., Hammack, C.A., Apicella, M.A. & Griffiss, J.M. (1988). Instability of expression of lipooligosaccharides and their epitopes in *Neisseria gonorrhoeae*. *Infect Immun* **56**: 942-946.

Schneider, H., Griffiss, J.M., Boslego, J.W., Hitchcock, P.J., Zahos, K.M. & Apicella, M.A. (1991). Expression of paragloboside-like lipooligosaccharides may be a necessary component of gonococcal pathogenesis in men. *Exp Med* **174**: 1601-1605.

Schryvers, A.B. & Stojiljkovic, I. (1999). Iron acquisition systems in the pathogenic *Neisseria*. *Mol Microbiol* **32**: 1117-1123.

Schwan, T.G., Schrumpf, M.E., Hinnebusch, B.J., Anderson, D.E.Jr. & Konkel, M.E. (1996). GlpQ: An antigen for serological discrimination between relapsing fever and Lyme Borreliosis. *J Clin Microbiol* **34**, 2483-2492.

Schwan, T.G. & Hinnebusch, B.J. (1998). Bloodstream- versus tick-associated variants of a relapsing fever bacterium. *Science* **280**: 1938-1940.

Schweda, E.K.H., Masoud, H., Martin, A., Risberg, A., Hood, D.W., Moxon, E.R., Weiser, J.N. & Richards, J.C. (1997). Phase variable expression and characteriation of phosphorylcholine oligosaccharide epitopes in *Haemophilus influenzae* lipopolysaccharides. *Glycoconj J* **14** (Suppl): S23.

Seifert, H.S., Ajioka, R.S., Marchal, C., Sparling, P.F. & So, M. (1988). DNA transformation leads to pilin antigenic variation in *Neisseria gonorrhoeae*. *Nature* **336**: 392-395.

Seiler, A., Reinhardt, R., Sarkari, J, Caugent, D.A. & Achtman, M. (1996). Allelic polymorphism and site-specific recombination in the *opc* locus of *Neisseria meningitidis*. *Mol Microbiol*. **19**: 841-856.

Shahin, R.D., Brennan, M.J., Li, Z.M., Meade, B.D. & Manclark, C.R. (1990). Characterization of the protective capacity and immunogenicity of the 69-kD outer membrane protein of *Bordetella pertussis*. *J Exp Med* **171**: 63-73.

Shevchenko, D.V., Akins, D.R., Robinson, E.J., Li, M., Shevchenko, O.V. & Radolf, J.D. (1997). Identificaton of homologues for thioredoxin, peptidyl prolyl cis isolmerase, and glycerophosphodiester phosphodiesterase in outer membrane fractions from *Treponema pallidum*, the syphilis spirochete. *Infect Immun* **65**, 4179-4189.

Silverblatt, F.J. (1974). Host-parasite interaction in the rat renal pelvis: a possible role for pili in the pathogenesis of pyelonephritis. *J. Exp Med* **140**: 1696-1711.

Silverblatt, F.J. & Ofek, J. (1978). Influence of pili on the virulence of *Proteus mirabilis* in experimental hematogenous pyelonephritis. *J Infect Dis* **138**: 664-667.

Silverman, M., Zieg, J. & Simon, M. (1979). Flagellar-phase variation: isolation of the *rh1* gene. *J Bacteriol* **137**: 517-523.

Silverman, M. & Simon, M. (1980). Phase variation: genetic analysis of switching mutants. *Cell* **19**: 845-854.

Simon, D. & Rest, R.F. (1992). *Escherichia coli* expressing a *Neisseria gonorrhoeae* opacity-associated outer membrane protein invade human cervical and endometrial epithelial cell lines. *Proc Natl Acad Sci USA* **89**: 5512-5516.

Singh, A.E. & Romanowski, B. (1999). Syphilis: Review with emphasis on clinical, epidemiologic, and some biologic features. *Clin Microbiol Rev* **12**: 187-209.

Skjak-Braek, G., Zanetti, F. & Paoletti, S. (1989). Effect of acetylation on some solution and gelling properties of alginates. *Carbohydrate Res* **185:** 131-138.

Smith, H. (1991). The Leeuwenhoek Lecture, 1991. The influence of the host on microbes that cause disease. *Proc R Soc Lond B* **246**: 97-105.

Snellings, N.J., Johnson, E.M. Kopecko, D.J., Collins, H.H. & Baron, L.S. (1981). Genetic regulation of variable Vi antigen expression in a strain of *Citrobacter freundii*. *J Bacteriol* **145**: 1010-1017.

Sparling, P., Cannon, J. & So, M. (1986). Phase and antigenic variation of pili and outer membrane protein II of *Neisseria gonorrhoeae*. *J Infect Dis* **153**: 196-201.

Spratt, B.G., Bowler, L.D., Zhang, Q.Y., Zhou, J. & Smith, J.M. (1992). Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal *Neisseria* species. *J Mol Evol* **34**: 115-125.

Stebeck, C.E., Shaffer, J.M., Arroll, T.W., Lukehart, S.A. & van Voorhis, W.C. (1997). Identification of the *Treponema pallidum* subsp. *pallidum* glycerophosphodiester phosphodiesterase homologue. *FEMS Micro Letts* **154**, 303-320.

Stefano, J.E. & Gralla, J.D. (1982). Spacer mutations in the *lac* ps promoter. *Proc Natl Acad Sci USA* **79**: 1069-1072.

Stephens, D.S. & McGee, Z.A. (1981). Attachment of *Neisseria meningitidis* to human mucosal surfaces: influence of pili and type of receptor cell. *J Infect Dis* **143**: 525-532.

Stephens, D.S., Spellman, P.A. & Swartley, J.S. (1993). Effect of the (α2→8)-linked polysialic acid capsule on adherence of *Neisseria meningitidis* to human mucosal cells. *J Infect Dis* **167**: 475-479.

Stern, A., Nickel, P., Meyer, T. & So, M. (1984). Opacity determinants of *Neisseria gonorrhoeae*: gene expression and chromosomal linkage to the gonoccocal pilus gene. *Cell* **37**: 447-456.

Stern, A., Brown, M., Nickel, P. & Meyer, T. (1986). Opacity genes in *Neisseria gonorrhoeae*: control of phase and antigenic variation. *Cell* **47**: 61-71.

Stern, A. & Meyer, T.F. (1987). Common mechanism controlling phase and antigenic variation in pathogenic *Neisseria*. *Mol Microbiol* **1**: 5-12.

Sternbach, H., Engelhardt, R. & Lezius, A.G. (1975). Rapid isolation of highly active RNA polymerase from *Escherichia coli* and its subunits by matrix-bound heparin. *Eur J Biochem* **60**: 51-55.

Stewart, F.M. (1994). Fluctuation tests: how reliable are the estimates of mutation rates? *Genetics* **137**: 1139-1146.

Stewart, F.M., Gordon, D.M. & Levin, B.R. (1990). Fluctuation analysis: the probability distribution of the number of mutants under different conditions. *Genetics* **124**: 175-185.

Stibitz, S., Aaronson, W., Monack, D. & Falkow, S. (1988). The *vir* locus and phase-variation in *Bordetella pertussis*. *Tokai J Exp Clin Med* **13**: Suppl 223-226.

Stibitz, S., Aaronson, W., Monack, D. & Falkow, S. (1989). Phase variation in *Bordetella pertussis* by frameshift mutation in a gene for a novel two-component system. *Nature* **338**, 266-269.

Stimson, E., Virji, M., Makepeace, K., Dell, A., Morris, H.R., Payne, G., Saunders, J.R., Jennings, M.P., Barker, S., Pancino, M., Blench, I. & Moxon, E.R. (1995). Meningococcal pilin: a glycoprotein substituted with digalactosyl 2,4-diacetamido-2,4,6-trideoxyhexose. *Mol Microbiol* **17**: 1201-1214.

Stimson, E., Virji, M., Barker, S., Panico, M., Blench, I., Saunder, J., Payne, G., Moxon, E.R., Dell, A. & Morris, H.R. (1996). Discovery of a novel protein modification: alpha-glycerophosphate is a substituent of meningococcal pilin. *Biochem J* **316**: 29-33.

Streisinger, G. & Owen, J. (1985). Mechanisms of spontaneous and induced frameshift mutation in bacteriophage T4. *Genetics* **109**: 633-659.

Stocker, B.A.D. (1949). Measurements of rate of mutation of flagellar antigenic phase in *Salmonella typhimurium*. *J Hygiene* **47**: 398-413.

Stoenner, H.G., Dodd, T. & Larsen, C. (1982). Antigenic variation in *B. hermsii*. *J Exp Med* **156**: 1297-1311.

Strauss, E.J. & Falkow, S. (1997). Microbial pathogenesis: genomics and beyond. *Science* **276**: 707-712.

Strzelecka, T.E., Clore, G.M. & Gronenborn, A.M. (1995). The solution structure of the Mu Ner protein reveals a helix-turn-helix DNA recognition motif. *Structure* **3**: 1087-1095.

Sugasawara, R.J., Cannon, J.G., Black, W.J., Nachamkin, I., Sweet, R.L. & Brooks, G.F. (1983). Inhibition of *Neisseria gonorrhoeae* attachmnet to HeLa cells with monoclonal antibody directed against a protein II. *Infect Immun* **42**: 980-985.

Swanson, J. (1978). Studies on gonococcus infection. XII. Colony color and opacity variants of gonococci. *Infect Immun* **19**: 320-331.

Swanson, J., Bergstrom, S., Robbins, K., Barrera, O., Corwin, D. & Koomey, J.M. (1986). Gene conversion involving the pilin structural gene correlates with pilus$^+$ (in equilibrium with) pilus$^-$ changes in *Neisseria gonorrhoeae*. *Cell* **47**: 267-276.

Swanson, J., Morrison, S., Barrera, O. & Hill, S. (1990). Piliation changes in transformation-defective gonococci. *J Exp Med* **171**: 2131-2139.

Szabo, M., Maskell, D., Butler, P., Love, J. & Moxon, R. (1992). Use of chromosomal gene fusions to investigate the role of repetitive DNA in regulation of genes involved in lipopolysaccharide biosynthesis in *Haemophilus influenzae*. *J Bacteriol* **174**: 7245-7252.

Taddei, F., Matic, I., Godelle, B. & Radman, M. (1997). To be a mutator, or how pathogenic and commensal bacteria can evolve rapidly. *Trends Microbiol* **5**: 427-8.

Taha, M-K. (1993). Increased sensitivity of gonococcal *pilA* mutants to bactericidal activity of normal human serum. *Infect Immun* **61**: 4662-4668.

Taha, M-K., Giorgini, D. & Nassif, X. (1996). The *pilA* regulatory gene modulates the pilus-mediated adhesion of *Neisseria meningitidis* by controlling the transcription of *pilC1*. *Mol Microbiol* **19**: 1073-1084.

Taha, M-K., Morand, P., Pereira, Y., Eugene, E., Giorgini, D., Larribe, M. & Nassif, X. (1998). Pilus-mediated adhesion of *Neisseria meningitidis* – the essential role of cell contact-dependent transcripional upregulation of the PilC1 protein. *Mol Microbiol* **28**: 1153-1163.

Taylor, R.K., Miller, V.L/. Furlong, D.B. & Mekalanos, J.J. (1987). Use of *phoA* gene fusions to identify a pilus colonization factor coordinately regulated with cholera toxin. *Proc Natl Acad Sci USA* **84**: 2833-2837.

Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., Ketchum, K.A., Hood, D.W., Peden, J.F., Dodson, R.J., *et al.* (2000). Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* 278: 1809-1815.

Theiss, P.M., Kim, M.F. & Wise, K.S. (1993). Differential protein expression and surface presentation generate high-frequency antigenic variation in *Mycoplasma fermentans*. *Infect Immun* **61**: 5123-5128.

Theiss, P. & Wise, K.S. (1997). Localized frameshift mutation generates selective, high-frequency phase variation of a surface lipoprotein encoded by a Mycoplasma ABC transporter operon. *J Bacteriol* **179**: 4013-4022.

Thony-Meyer, L., Beck, C., Preisig, O. & Hennecke, H. (1994). The *ccoNOQP* gene cluster codes for a *cb*-type cytochrome oxidase that functions in aerobic respiration of *Rhodobacter capsulatus*. *Mol Microbiol* **14**: 705-716.

Tomb, J-F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., *et al.* (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**: 539-547.

Tommassen, J., Vermeij, P., Struyve, M., Benz, R. & Poolman, J.T. (1990). Isolation of *Neisseria meningitidis* mutants deficient in class 1 (PorA) and class 3 (PorB) outer membrane proteins. *Infect Immun* **58**: 1355-1359.

Tramont, E.C. (1995). Treponema pallidum (syphilis). pages: 2117-2133. in: Mandell, Douglas and Bennett's Principles and Practice of Infectious Diseases (Fourth Edition), Eds: Mandell, G.L., Bennett, J.E. & Donlin, R. Churchill Livingstone.

Tran, H.T., Keen, J.D., Kricker, M., Resnick, M.A. & Gordenin, D.A. (1997). Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. *Mol Cell Biol* **17**, 2859-2865.

Tsai, C-M., Frasch, C.E. & Mocca, L.F. (1981). Five structural classes of major outer membrane proteins in *Neisseria meningitidis*. *J Bacteriol* **146**: 69-78.

Tuomanen, E. & Weiss, A. (1985). Characterization of two adhesins of *Bordetella pertussis* for human ciliated respiratory-epithelial cells. *J Infect Dis* **152**: 118-125.

Tuomanen, E., Towbin, H., Rosenfelder, G., Braun, D., Larson, G., Hansson, G.C. & Hill, R. (1988). Receptor analogs and monoclonal antibodies that inhibit adherence of *Bordetella pertussis* to human ciliated respiratory epithelial cells. *J Exp Med* **168**: 267-277.

Urisu, A., Cowell, J.L. & Manclark, C.R. (1986). Filamentous hemagglutinin has a major role in mediating adherence of *Bordetella pertussis* to human WiDr cells. *Infect Immun* **52**: 695-701.

van Belkum, A., Scherer, S., van Leeuwen, W., Willemse, D., van Alphen, L. & Verbrugh, H. (1997a). Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*. *Infect Immun* **65**: 5017-5027.

van Belkum, A., Melchers, W.J.G., Ijsseldijk, C., Nohlmans, L., Verbuch, H. & Meis, J.F.G.M. (1997b). Outbreak of amoxycillin-resistant *Haemophilus influenzae* type b: variable number of tandem repeats as novel molecular markers. *J Clin Microbiol* **35**: 1517-1520.

van der Ley (1988). Three copies of a single protein II-encoding sequence in the genome of *Neisseria gonorrhoeae* JS3: evidence for gene conversion and gene duplication. *Mol Microbiol* **2**: 797-806.

van der Ley, P., van der Biezen, J., Sutmuller, R., Hoogerhout, P. & Poolman, J.T. (1996). Sequence variability of FrpB, a major iron-regulated outer-membrane protin in the pathogenic neisseriae. *Microbiology* **142**: 3269-3274.

van Ham, S.M., van Alphen, L., Mooi, F.R., van Putten, J.P.M. (1993). Phase variation of *H. influenzae* fimbriae: transcriptional control of two divergent genes through a variable combined promoter region. *Cell* **73**: 1187-1196.

van Putten, J.P.M. (1993). Phase variation of lipopolysaccharide directs interconversion of invasive and immuno-resistant phenotypes of *Neisseria gonorrhoeae*. *EMBO J* **12**: 4043-4051.

van Putten, J.P.M. & Robertson, B.D. (1995) Molecular mechanisms and implication for infection of lipopolysaccharide variation in *Neisseria*. *Mol Microbiol* **16**: 847-853.

van den Akker, W.M.R. (1998). Lipopolysaccharide expression within the genus *Bordetella*: influence of temperature and phase variation. *Microbiology* **144**: 1527-1535.

van der Ende, A., Hopman, C.T.P., Zaat, S., Oude Essink, B.B., Berkhout, B. & Dankert, J. (1995). Variable expression of class 1 outer membrane protein in *Neisseria meningitidis* is caused by variation in the spacing between the –10 and –35 regions of the promoter. *J Bacteriol* **177**: 2475-2480.

van der Ley, P. (1988). Three copies of a single protein II-encoding sequence in the genome of *Neisseria gonorrhoeae* JS3: evidence for gene conversion and gene duplication. *Mol Microbiol* **2**: 797-806.

van der Ley, P., van der Biezen, J., Sutmuller, R., Hoogerhout, P. & Poolman, J.T. (1996). Sequence variability of FrpB, a major iron-regulated outer-membrane protin in the pathogenic neisseriae. *Microbiology* **142**: 3269-3274.

van der Woulde, M.W., Braaten, B.A. & Low, D.A. (1992). Evidence for global regulatory control of pilus expression in *Escherichia coli* by Lrp and Dam methylation: model based on analysis of *pap*. *Mol Microbiol* 6: 2429-2435.

Van Dop, C., Yamanaka, G., Steinberg, F., Sekura, R.D., Manclark, C.R., Stryer, L. & Bourne, H.R. (1984). ADP ribosylation of transducin by pertussis toxin blocks the light-stimulated hydrolysis of GTP and cGMP in retinal photoreceptors. *J Biol Chem* **259**: 23-26.

Vari, F. & Bell, K. (1996). A simplified silver diammine method for the staining of nucleic acids in polyacrylamide gels. *Electrophoresis* **17**: 20-25.

Vazquez, J.A., Berron, S., O'Rourke, M., Carpenter, G., Feil, E., Smith, N.H. & Spratt, B.G. (1995). Interspecies recombination in nature: a meningococcus that has acquired a gonococcal PIB porin. *Mol Microbiol* **15**: 1001-1007.

Virji, M., Weiser, J.N., Lindberg, A.A. & Moxon, E.R. (1990). Antigenic similarities in lipopolysaccharides of *Haemophilus* and *Neisseria* and expression of a digalactoside structure also present on human cells. *Microb Pathog* **9**: 441-450.

Virji, M., Kayhty, H., Fergusson, D.J.P., Alexandrescu, C., Heckles, J.E. & Moxon, E.R. (1991). The role of pilin in the interactions of pathogenic *Neisseria* with cultured human endothelial cells. *Mol Microbiol* **5**: 1831-1841.

Virji, M., Makepeace, K., Ferguson, D.J., Achtman, M., Sarkari, J. & Moxon, E.R. (1992a). Expression of the Opc protein correlates with invasion of epithelial and endothelial cells by *Neisseria meningitidis*. *Mol Microbiol* **6**: 2785-2795.

Virji, M., Alexandrescu, C., Fergusson, D.J.P., Saunders, J.R. & Moxon, E.R. (1992b). Variations in the expression of pili: the effect on adherence of *Neisseria meningitidis* to human epithelial and endothelial cells. *Mol Microbiol* **6**: 1271-1279.

Virji, M., Makepeace, K., Ferguson, D.J., Achtman, M. & Moxon, E.R. (1993a). Menigococcal Opa and Opc proteins: their role in colonisation and invasion of human epithelial and endothelial cells. *Mol Microbiol* **10**: 499-510.

Virji, M., Saunders, J.R., Sims, G., Makepeace, K., Maskell, D. & Ferguson, D.J.P. (1993b). Pilus-facilitated adherence of *Neisseria meningitidis* to human epithelial and endothelial cells: modulation of adherence phenotype occurs concurrently with changes in the primary amino acid sequence and the glycosylation status of pilin. *Mol Microbiol* **10**: 1013-1028.

Virji, M., Makepeace, K. & Moxon, E.R. (1994). Distinct mechanisms of interactions of Opc-expressing meningococci at apical and basolateral surfaces of human endothelial cells; the role of integrins in apical interactions. *Mol Microbiol* **14**: 173-184.

Virji, M., Makepeace, K., Peak, I.R.A., Ferguson, D.J.P., Jennings, M.P. & Moxon, E.R. (1995). Opc- and pilus-dependent interactions of meningococci with human endothelial cells: molecular mechanisms and modulation by surface polysaccharides. *Mol Microbiol* **18**: 741-754.

von Hippel, P.H., Bear, D.G., Morgan, W.D. & McSwiggen, S.A. (1984). Protein-nucleic acid interactions in transcription: a molecular analysis. *Annu Rev Biochem* **53**: 389-446.

Wainwright, L.A., Pritchard, K.H. & Seifert, H.S. (1994). A conserved DNA sequence is required for efficient gonococcal pilin antigenic variation. *Mol Microbiol* **13**: 75-87.

Waldbeser, L.S., Ajioka, R.S., Merz, A.J., Puaoi, D., Lin, L., Thomas, M. & So, M. (1994). The OpaH locus of *Neisseria gonorrhoeae* MS11A is involved in epithelial cell invasion. *Mol Microbiol* **13**: 919-928.

Walker, E.M., Borenstein, L.A., Blanco, D.R., Miller, J.N. & Lovell, M.A. (1991). Analysis of outer membrane ultrastructure of pathogenic *Treponema* spp. *Borrelia* spp. and by freeze-fracture electron microscopy. *J Bacteriol* **173**, 5585-5588.

Warne, S.E. & deHaseth, P.L. (1993). Promoter recognition by *Escherichia coli* RNA polymerase. Effects of single base pair deletions and insertions in the spacer DNA separating the –10 and –35 regions are dependent on spacer DNA sequence. *Biochemistry* **32**: 6134-6140.

Weel, J.F.L., Hopman, C.T.P. & van Putten, J.P.M. (1989). Stable expression of lipooligosaccharide antigens during attachment, internalization, and intracellular processing of *Neisseria gonorrhoeae* in infected epithelial cells. *Infect Immun* **57**: 3395-3402.

Weel, J.F.L. & van Putten, J.P.M. (1991). Fate of the major outer membrane protein P.IA in early and late events of gonococcal infection of epithelial cells. *Res Microbiol* **142**: 985-993.

Weel, J.F.L., Hopman, C.T.P. & van Putten, J.P.M. (1991a). *In situ* expression and localization of *Neisseria gonorrhoeae* opacity proteins in infected epithelial cells: apparent role of Opa proteins in cellular invasion. *J Exp Med* **173**: 1395-1405.

Weel, J.F.L., Hopman, C.T.P. & van Putten, J.P.M. (1991b). Bacterial entry and intracellular processing of *Neisseria gonorrhoeae* in epithelial cells: immunomorphological evidence for alterations in the major outer membrane protein P.1B. *J Exp Med* **174**: 705-717.

Weinberg, E.D. (1978). Iron and infection. *Microbiol Rev* **42**: 45-66.

Weiser J.N., Lindberg, A.A., Manning, E.J., Hansen, E.J. & Moxon, E.R. (1989a). Identification of a chromosomal locus for expression of lipopolysaccharide epitopes in *Haemophilus influenzae*. Infect Immun **57**: 3945-3052.

Weiser, J.N., Love, J.M. & Moxon, E.R. (1989b). The molecular mechanism of phase variation of *H. influenzae* lipopolysaccharide. *Cell* **59**: 657-665.

Weiser, J.N., Maskell, D.J., Butler, P.D., Lindberg, A.A. & Moxon, E.R. (1990). Characterisation of repetitive sequences controlling phase variation in *Haemophilus influenzae* lipopolysaccharide. *J Bacteriol* **172**: 3304-3309.

Weiser, J.N. (1993). Relationship between colony morphology and the life cycle of *Haemophilus influenzae:* The contribution of lipopolysaccharide phase variation to pathogenesis. *J Infect Dis* **168**: 672-680.

Weiser, J.N., Chong, S.T.H., Greenberg, D. & Fong, W. (1995). Identification and characterisation of a cell envelope protein of *Haemophilus influenzae* contributing to phase variation in colony opacity and nasopharyngeal colonization. *Mol Microbiol* **17**: 555-564.

Weiser, J.N., Shchepetov, M. & Chong, S.T.H. (1997). Decoration of lipopolysaccharide with phosphorylcholine: a phase-variable chracteristic of *Haemophilus influenzae. Infect Immun* **65**: 943-950.

Weiser, J.N. & Pan, N. (1998). Adaptation of *Haemophilus influenzae* to acquired and innate humoral immunity based on phase variation of lipopolysaccharide. *Mol Microbiol* **30**: 767-775.

Weiser, J.N., Pan, N., McGowan, K.L., Musher, D., Martin, A. & Richards, J. (1998a). Phosphorylcholine on the lipopolysaccharide of *Haemophilus influenzae* contributes to persistance in the respiratory tract and sensitivity to serum killing mediated by C-reactive protein. *J Exp Med* **187**: 631-640.

Weiser, J.N., Goldberg, J.B., Pan, N., Wilson, L. & Virji, M. (1998b) The phosphoylcholine epitope undergoes phase variation on a 43-kilodalton protein in *Pseudomonas aeruginosa* and on pili of *Neisseria meningitidis* and *Neisseria gonorrhoeae. Infect Immun* **66**, 4263-4267.

Weiss, A.A., Hewlett, E.L., Myers, G.A. & Falkow, S. (1983). Tn5-induced mutations affecting virulence factors of *Bordetella pertussis. Infect Immun* **42**: 33-41.

Wiertz, E.J.H.J., Delvig, A., Donders, E.M.L.M., Brugghe, H.F., van Unen, L.M.A., Timmermans, H.A.M., Achtman, M., Hoogerhout, P. & Poolman, J.T. (1996). T-cell responses to outer membrane proteins of *Neisseria meningitidis*: comparative study of the Opa, Opc and PorA proteins. *Infect Immun* **64**: 298-304.

Willems, R., Paul, A., van der Heide, H.G.J., ter Avest, A.R. & Mooi, F.R. (1990). Fimbrial phase variation in *Bordetella pertussis*: a novel mechanism for transcriptional regulation. *EMBO J* **9**: 2803-2809.

Wise, K.S. (1993). Adaptive surface variation in mycoplasmas. *Trends Microbiol* **1**: 59-63.

Wise, K.S., Kim, M.F., Theiss, P.M. & Lo S-C. (1993). A family of strain-variant surface lipoproteins of *Mycoplasma fermentans*. *Infect Immun* **61**: 3327-3333.

Witkin, E.M. (1946). Inherited differences in sensitivity to radiation in *Escherichia coli*. *Proc Natl Acad Sci Wash* **32**: 59-68.

Wolfgang, M., Park, H-S., Hayes, S.F., van Putten, J.P.M. & Koomey, M. (1998). Suppression of an absolute defect in type IV pilus biogenesis by loss-of-function mutations in *pilT*, a twitching motility gene in *Neisseria gonorrhoeae*. *Proc Natl Acad Sci USA* **95**: 14973-14978.

Woods, J.P., Spinola, S.M., Strobel, S.M. & Cannon, J.G. (1989). Conserved lipoprotein H.8 of pathogenic *Neisseria* consists entirely of pentapeptide repeats. *Mol Microbiol* **3**: 43-48.

Woods, J.P. & Cannon, J.G. (1990). Variation in expression of class 1 and class 5 outer membrane proteins during nasopharyngeal carriage of *Neisseria meningitidis*. *Infect Immun* **58**: 569-572.

Wright, S.W., Edwards, K.M., Decker, M.D. & Zeldin, M.H. (1995). Pertussis infection in adults with persistent cough. *JAMA* **273**: 1044-1046.

Yang, Q.L. & Gotschlich, E.C. (1996). Variation of gonococcal lipooligosaccharide structure is die to alterations in poly-G tracts in *lgt* genes encoding glycosyl transferases. *J Exp Med* **183**: 323-327.

Yogev, D., Rosengarten, R., Watson-McKown, R. & Wise, K.S. (1991). Molecular basis of *Mycoplasma* surface antigenic variation: a novel set of divergent genes undergo spontaneous mutation of periodic coding regions and 5′ regulatory sequences. *EMBO J* **10**: 4069-4079.

Yogev, D., Menaker, D., Strutzberg, K., Levisohn, S., Kirchhoff, H., Hinz, K-H. & Rosengarten, R. (1994). A surface epitope undergoing high-frequency phase variation is shared by *Mycoplasma gallisepticum* and *Mycoplasma bovis*. *Infect Immun* **62**: 4962-4968.

Yogev, D., Watson-McKown, R., Rosenbgarten, R., Im, J. & Wise, K.S. (1995). Increased structural and combinatorial diversity in an extended family of genes encoding Vlp surface proteins of *Mycoplasma hyorhinis*. *J Bacteriol* **177**: 5636-5643.

Zhang, J.M., Cowell, J.L., Steven, A.C., Carter, P.H., McGrath, P.P. & Manclark, C.R. (1985). Purification and characterization of fimbriae isolated from *Bordetella pertussis*. *Infect Immun* **48**: 422-427.

Zhang, J.-R. & Norris, S.J. (1998). Genetic variation of the *Borrelia burgdorferi* gene *vlsE* involves cassette-specific, segmental gene conversion. *Infect Immun* **66**: 3698-3704.

Zhang, J.-R., Hardham, J.M., Barbour, A.G. & Norris, S.J. (1997). Antigenic variation in Lyme disease borreliae by promiscuous recombination of *vmp*-like sequence cassettes. *Cell* **89**: 275-285.

Zhang, Q. & Wise, K.S. (1996). Molecular basis of size and antigenic variation of a *Mysoplasma hominis* adhesin encoded by divergent *vaa* genes. *Infect Immun* **64**: 2737-2744.

Zhang, Q. & Wise, K.S. (1997). Localized reversible frameshift mutation in an adhesin gene confers a phase-variable adherence phenotype in mycoplasma. *Mol Microbiol* **25**: 849-869.

Zhang, Q.Y., DeRyckere, D., Lauer, P. & Koomey, M. (1992). Gene conversion in *Neisseria gonorrhoeae*: evidence for its role in pilus antigenic variation. *Proc Natl Acad Sci USA* **89**: 5366-5370.

Zhao, H., Li, X., Johnson, D.E., Blomfield, I. & Mobley, H.L. (1997). *In vivo* phase variation of MR/P fimbrial gene expression in *Proteus mirabilis* infecting the urinary tract. *Mol Microbiol* **23**: 1009-1019.

Zhou, J. & Spratt, B.G. (1992). Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*: interspecies recombination within the *argF* gene. *Mol Microbiol* **6**: 2135-2146.

Zhu, P., Morelli, G., Linz, B. & Achtman, M. (1998). Evolution of the *opc* region in the *Neisseriae*. In: Nassif, X., Quentin-Millet, M-J. & Taha, M-K (Eds). Abstracts of the eleventh international pathogenic *Neisseria* conference. E.D.K., Paris, 1998.

Zieg, J., Silverman, M., Hilmen, M. & Simon, M. (1977). Recombinational switch for gene expression. *Science* **196**: 170-172.

Zieg, J., Hilmen, M. & Simon, M. (1978). Regulation of gene expression by site-specific inversion. *Cell* **15**: 237-244.

Zieg, J. & Simon, M. (1980). Analysis of the nucleotide sequence of an invertible controlling element. *Proc Natl Acad Sci USA* **77**: 4196-4200.

Appendix 1

An analysis of DNA repeats in *Haemophilus influenzae* strain Rd.

Two analyses of the repeat sequences present in the complete genome sequence of *H. influenzae* strain Rd in order to identify phase variable genes have been published (Hood *et al.*, 1996; van Belkum *et al.*, 1997a). In the first analysis, a combined approach using FINDPATTERNS and BLASTN searches was used, in the second a repeat finding algorithm was used that looked for repeats without regard for the component motifs. The findings of the two studies are summarised in Tables 9.1 and 9.2.

Table 9.1: Summary of the results of Hood *et al.*'s analysis of the genome sequence of *H. influenzae* strain Rd.

Repeat motif	Number of repeats	Gene homology
AAT *	9	Heme-utilisation gene *hxuC*
CAAT	17	Lic1
CAAT	22	Lic2
CAAT	32	Lic3
GCAA	25	YadA
GACA	22	LgtC
CAAC	36	Hemoglobin receptor
CAAC	20	Hemoglobin receptor
CAAC	18	Hemoglobin receptor
CAAC	20	Hemoglobin receptor
CAAC	15	None
AGTC	32	Methyltransferase
TTTA	6	32.9-Kda protein of unknown function

* This repeat is located in the promoter region of the associated ORF and the authors did not indicate that they believe it to affect gene expression, although they suggest that it might.

Note: This analysis did include searches for homopolymeric tracts and dinucleotide repeats. Since the number of longer homopolymeric tracts did not differ from predictions based upon %GC and random base distribution, and the longer dinucleotide repeats were not uniquely within, or associated with, ORFs – neither set of search results were considered to be markers for phase variable genes.

Table 9.2: Summary of the results of van Belkum *et al.*'s analysis of the genome sequence of *H. influenzae* strain Rd.

Repeat code	Repeat motif	Number of repeats	Gene homology*	Evidence for variation
2-1	AT	5	NA	NA
2-2	GC	5	NA	NA
2-3	AC	5	NA	NA
2-4	TG	5	NA	NA
2-5	TC	5	NA	NA
3-1	ATT	9	NA	+
4-1	GTCT	22	LgtC	+
4-2	CAAT	32	Lic3	+
4-3	CAAT	23	Lic2	+
4-4	TTGG	21	Hemoglobin receptor	+
4-5	TTGG	20	Hemoglobin receptor	+
4-6	TTTA	6	32.9-kDA protein	-
4-7	TTGG	37	Hemoglobin receptor	+
4-8	TGAC	20	Methyltransferase	+
4-9	TTGG	16	No homology	+
4-10	TTGC	25	YadA	+
4-11	CAAT	17	Lic1	+
4-12	TTGG	19	Hemoglobin receptor	+
5-1	TTATC	12	NA	+**
5-2	GTCTC	4	NA	+
6-1	CTGGCT	4	NA	+
6-2	GGCAAT	3	NA	+
6-3	TTAAAA	3	NA	-

* These homologies were reported by van Belkun *et al.*, on the basis of the previous paper by Hood *et al.*, and not on the basis of their own analysis. They are included here to aid comparison of search results.

** Amplification experiment suggests variation but there are multiple PCR products in the presented gel.

NA = not addressed.

The descriptions of the repeat motifs and the number of repetitions are not concordant (especially of the AGTC/TGAC associated methyltransferase). In order to resolve these differences, to determine whether the newly developed system could detect any novel repeat associated genes, and to identify the homologies and potential function of the genes identified in the second study that were not found in the first, a single pass analysis of the

H. influenzae strain Rd genome was performed using the ACEDB graphicial interface described in chapters 4, 5 and 6. This analysis has been performed only once, the results have not been re-searched using tabulated ARRAY and TANDEM results and the results of this analysis should not be considered exhaustive. However, the search was not directed by the results of previous reports and no previously reported repeat associated gene was missed. The search parameters used were for homopolymeric tracts of greater than 5 bases, dinucleotide repeats with 4 or more copies, and 3 or more copies of all repeats with motifs of 4 to 10 bases in length using ARRAY. The results of the search are summarised in Table 9.3.

Table 9.3: Results of a 'single pass' analysis of the *H. influenzae* strain Rd sequence.

Repeat identifier	Repeat motif*	Number of repeats	Gene homology / putative function	vB	H
1-A	A	8	Peptidoglycan biosynthesis	NA	-
1-B	T	8	Sensor / regulator protein	NA	-
1-C	T	8	Surface protein (*troC & troD*)[1]	NA	-
1-D	A	8	Peptidoglycan biosynthesis	NA	-
1-E	T	9	same ORF as 1-D	NA	-
1-F	T	7	LPS biosynthesis / galactosyltransferase	NA	-
1-G	T	8	Unknown function	NA	-
1-H	A	8	Membrane protein – same ORF as 5-B	NA	-
1-I	A	10	Iron binding protein[2]	NA	-
1-J	G	8	Transcriptional regulator / repressor	NA	-
1-K	G	8	Hypothetical	NA	-
1-L	T	9	Surface adhesin[3]	NA	-
1-M	A	9	Membrane protein	NA	-
1-N	A	8	Unknown function	NA	-
1-O	A	8	Hypothetical	NA	-
2-A	AT	5	Formate dehydrogenase	2-1	-
2-B	TA	4	Cytosine specific methyltransferase	-	-
3-A	ATT	9	Iron binding protein	3-1	+
4-A	AGAC	22	LPS biosynthesis / LgtC[4]	4-1	+
4-B	GCAA	3	ABC transporter	-	-
4-C	TCAA	33	LPS biosynthesis / Lic3	4-2	+
4-D	ATCA	23	LPS biosynthesis / Lic2	4-3	+
4-E	CCAA	21	Iron binding protein	4-4	+
4-F	CCAA	20	Iron binding protein	4-5	+

4-G	TTTA	6	Unknown function protein	4-6	+
4-H	TGTT	3	Unknown function protein	-	-
4-I	CCAA	37	Iron binding protein	4-7	+
4-J	TGCT	3	In same ORF as 4-I	-	-
4-K	CAAA	3	Transcriptional activator	-	-
4-L	CAGA	3	Unknown function	-	-
4-M	TATT	3	Potassium transport protein KefC	-	-
4-N	TATT	3	Unknown function	-	-
4-O	AGTC	32	Type III restriction / modification system	4-8	-
4-P	ATAA	3	Surface adhesin[5]	-	-
4-Q	ATCA	3	Unknown function	-	-
4-R	CTTT	4	Unknown function	-	-
4-S	AATT	3	DNA helicase	-	-
4-T	ACTC	3	DNA gyrase	-	-
4-U	AACC	3	ABC transporter	-	-
4-V	GTTT	3	Phosphate permease	-	-
4-W	AACC	16	Hypothetical	4-9	+
4-X	TAAT	3	Would not alter expression[5]	-	-
4-Y	GCAA	25	Adhesin (*YadA* or *Yop1* from *Yersinia*)	4-10	+
4-Z	CAAT	17	LPS biosynthesis / Lic1	4-11	+
4-AA	CCAA	19	Iron binding protein	4-12	+
4-BB	TATT	3	Would not alter expression	-	-
5-A	TTATC	12	Hypothetical / None [6]	5-1	NA
5-B	TTATT	3	Membrane protein	-	NA
5-C	TTACC	3	Peptidoglycan biosynthesis	-	NA
5-D	TCGTC	4	Methyltransferase	5-2	NA
6-A	CGTTTA	3	Unknown function	-	NA
6-B	AGCCAG	4	Hypothetical	6-1	NA
6-C	TGGCAA	3	Surface adhesin	6-2	NA
6-D	AATTTT	3	Phosphoglycerate kinase[7]	6-3	NA
6-E	ATGGTA	3	Outer membrane protein[8]	-	-
9-A	CGCCTTGTT	4	Hypothetical	NA	NA

* Motifs are described in the orientation of the associated open reading frame and start at the first base of the repeat element. vB indicates results from the van Belkum analysis, H indicates results from the Hood analysis. NA = not addressed, numerical designations are those reported in the original paper, + indicates previously reported, - indicates not previously reported.

Notes:

1. Repeat is associated with a frame-shift.
2. This repeat is in the promoter region of an ORF with homology with transferrin binding protein. There are several possible –10 / TATAAT sites (including perfect versions) 5′ of the initiation codon.

However, none leave space for a ribosomal binding site. There is a GATAAT just 3′ of the repeat. If this is the functional –10 region then alteration in repeat length would alter the relative position of the –35 and any other upstream promoter components.

3. 4-P and 1-L occur together as (ATAA)3-TC-(T)9 and the C of this sequence lies in the probable –35 position in the promoter. Instability in the (T)9 would affect positioning in the region of the –35 and alterations in the (ATAA)3 repeat, which in this sequence extends for one helical turn, would affect the position of any 3′ promoter components.

4. The strain Rd sequence analysed includes two frame-shifts that are not associated with the repeat suggesting that it is a dead gene. Subsequent sequencing of this locus has demonstrated that an intact gene is present (D. Hood – personal communication).

5. The final AT of this repeat are the first two bases of the ATG / initiation codon of the ORF. The repeat would lie between the promoter and the reading frame and the length of the repeat is unlikely to affect the phenotype.

6. In reverse the repeat includes termination codons in all frames. In the stated orientation there are termination codons in all frames – from this sequence this appears to be a 'dead' ORF.

7. This repeat is located 5′ of an ORF and there is a putative ribosomal binding site 3′ of the repeat. However, there is a perfect –10 / TATAAT 5′ of the repeat which would place the repeat in the intervening region between –10 and initiation codon. If this interpretation is correct, alteration in repeat length would not be expected to affect the expression of this gene.

8. This repeat is located at the 3′ end of the ORF. Changes in length would not be expected to alter expression. Changes in length might alter the antigenicity of the altered protein.

None of the dinucleotide repeats containing G or C bases were considered to be located appropriately or to be of sufficient length to mediate phase variation.

The search parameters used by van Belkum *et al.*, (1997a) to identify and describe the repeat elements are not explained in the paper. The predisposition to slippage within a repeat is likely to be a reflection of the number of component motifs. Even if the rate of variation is not simply a function of length, the number of whole motifs that constitute the repeat element might be expected to determine when a repeat becomes unstable. van Belkum *et al.*, reported the number of repeats detected as whole numbers but did not

search on this basis. For example, repeat elements composed of 6 mers were reported when the repeat is 20 bases in length (i.e. 3.3 copies) which were recorded as 3 copy repeats. However, the $(CGTTTA)_3$ repeat was not detected although it is, on the basis of repeat motif number, equally likely to be unstable.

The number of repeats reported in the tetramer-associated methyltransferase was correct in the report of Hood *et al.*, and in error in the report of van Belkum *et al.*,. The other differences in reported repeat numbers are related to the (unstated) conventions adopted by each author. Both groups chose to describe repeat motifs such that repeats with common components would be similarly desribed, rather than on the basis of the first base of each repeat sequence. Hood *et al.*, described the length of each repeat element on the basis of the number of complete copies of the repeat used to describe the element. In contrast, van Belkum *et al.* described repeats on the basis of the maximum number of component tetramers regardless of the starting base. In this sense the numbers reported by van Belkum *et al.*, more accurately describe the total number of repeated motifs in the repeat elements.

The differences between the motifs that were described between the studies is likewise due to differences in the (unstated) descriptive approaches adopted. Hood *et al.*, described repeats on the basis of the transcriptional orientation of the putative associated reading frames. van Belkum used the first motif of each type identified on the basis of the orientation in which the sequence had been loaded into the analysis system. They are therefore neither described on the basis of the orientation of the associated reading frames, nor on the basis of the directions of DNA replication – a rational alternative. It should be noted that the gene locations described by van Belkum are not, as stated, as in the published version of the *H. influenzae* strain Rd sequence. They are actually locations on the pre-publication sequence that was available from the TIGR ftp site prior to publication and cannot be directly related to the final version of the published sequence.

The component motifs of the tetrameric repeats identified in the *H. influenzae* strain Rd sequence fall into two categories. The first can be considered (on the basis of

transcriptional orientation) to be variants of 5′-CAAT-3′. Each repeat is a match for this sequence for at least 3 out of 4 bases and all have either CAA or AAT components. The second group of repeats can be described as 'G-C palindromes'. This description indicates that the repeats contain both C and G bases on each strand in such a way that they would have a predisposition to form C:G mediated intra-strand secondary structures. Some sequences (e.g. GCAA) can be considered to fall into both categories. The presence of tetrameric repeats is characteristic of *Haemophilus* spp.. This may simply be a reflection of which repeats occurred by chance in the species and were then distributed (with minor variations) within and between reading frames and strains. Alternatively, there may be species-specific processes which favour amplification or instability in repeats with these component motifs in *H. influenzae*.

The analysis described by Hood *et al.*, concentrated upon detecting and describing the strongest candidate phase variable genes present in the analysed genome. This was undoubtedly the most appropriate study that could be performed at that time. The analysis presented in this section has described many more repeats and it is unlikely that they will all be associated with phase variation. Based upon the repeats associated with phase variation in other species and those seen in other whole genome analyses, and their contextual locations, the strongest new candidates are: 1-C, 1-D/E, 1-I, 1-J, 1-K, 1-L, 1-M, 2-B, 3-A, 4-K, 4-L, 4-P & 4R. These 13 genes include homologues of one methyltransferase, one peptidoglycan biosynthesis gene, two iron binding proteins, two transcriptional regulators, two adhesins, two other surface / membrane proteins and three reading frames with no identifiable homologies. These functions are consistent with the functions of other phase variable genes previously identified in *H. influenzae* and other species.

The experimental work that accompanies the analysis of van Baelkum *et al.*, provides a number of important insights into the behaviour of repeat elements that are not addressed by the authors. The relative contributions of the repeat itself and the context in which it

exists to local instability have not been determined. Likewise the length of / number of iterations of the motif that compose a repeat that are required for instability is not known. van Belkum *et al.*, prepared primers and looked at repeat-associated polymorphisms in a collection of unrelated strains and also in a series of epidemiologically linked isolates. Instability was identified in hexameric repeats with only 3 and 4 copies (6-1 and 6-2). The shorter of these repeats contained the familiar CAAT motif. Likewise, some tetrameric repeats (especially 4-8, associated with a methyltransferase) were present in as few as two copies in some strains. It is therefore meaningful to include shorter repeat elements in this type of search for repeat-associated putative phase variable genes. There are 256 possible tetrameric repeat motifs and 64 motifs if their common components are described. Only 10% of these are detected in the *H. influenzae* strain Rd sequence, frequently in multiple locations, whilst the others are not present as repeats at all. Some of these would be selected against because they contain termination codons but there is the suggestion that in *H. influenzae*, repeats of this type are, except where they are associated with functional instability, less frequently present in the genome than would be expected by chance.

It would be expected that the rate of variation mediated by shorter repeats would be lower than by longer repeat elements. It is also possible repeats are lost during culture *in-vitro*. However, it is possible that these observations, revealing the presence of these shorter repeat elements in genes where similar repeats in other strains are believed to be functional, highlight the importance not only of the rate of variation with which longer tracts are associated - but also the type of mutation. The presence of a repeat may make a region relatively unstable due to a tendency to slippage, duplication and deletion. Even when this rate is very low, as it might be with very short repeat elements, it may still be higher than that in the surrounding sequence and act as a reversible switch. Furthermore, under conditions where the associated variation confers a fitness advantage, selection for variability, and thus for longer repeat elements, would lead to an increase in the repeat element lengths in a population exposed to appropriate changing environments.

The appropriate rate of variation for each phenotype to provide a fitness advantage may vary between genes and in some cases (possibly methyltransferases) may be lower than others. The function of phase variable methyltransferases is not known. If they are to generate a sub-population that have resistance to bacteriophages that are encountered infrequently, then the generation of variants at a high frequency would be unnecessary. If expression of these genes reduces fitness in other contexts, then a low rate of variation would be most adaptive. The tetramer associated with a methyltransferase is one of the few that does not include a 5′-CAAT-3′ related repeat and it is frequently present with as few as 2 or 3 repeated motifs (this gene is associated with another atypical repeat, AGCC, in some strains (Hood *et al.*, 1996)). The pentamer-associated methyltransferase has only 4 copies of a repeat that is 60% GC (and would have a relatively high melting temperature) and does not include a 5′-CAAT-3′ related sequence. Finally, the newly identified repeat of (TA)5 in the 5′ region of a cytosine-specific methyltransferase might also be expected to vary at a low rate. Finally, there is another methlytransferase that is frame-shifted in the *H. influenzae* strain Rd genome which includes a (T)5 at the junction – this might also be variable at a low rate.

The rate of variation necessary to confer a selective advantage on a population need not be in the frequently reported 10^{-3} to 10^{-4} /cell/generation range. The influence of selection is discussed in section 3.2. Selection may account for the observed variability reported for hexameric repeats 6-B and 6C but not 6-D (van Belkum *et al.*, 1997a). Of these, 6-D has the lowest melting temperature and might have been expected to be most unstable. However, according to the new analysis, 6-D is the only hexameric repeat with 3 or more copies that does not have the potential to alter phenotypic characteristics. The detected variability might reflect the combined effects of instability and selection and that, in terms of the composition of a bacterial population, in the absence of selective advantage the repeats are effectively stable.

Appendix 2

Definitions for working with repetitive sequence

Repetitive sequence:

The generic term for sequence including repeated elements.

Simple sequence:

When a sequence contains a repetition of bases or groups of bases that occur more frequently than would be expected to occur by chance.

note: 'repetitive sequence' is a subset of 'simple sequence'

Repeat region:

When a sequence contains a repetition of an identifiable base or group of bases which can be considered to represent a motif.

Recurrent sequence:

Any repeated DNA motif occurring at a distance - in different genetic locations.

Direct repeat:

Repeats with the components in the same orientation.

Inverted repeat:

Repeats with the components in inverted orientation

Continuous repeat:

Repeats in which there is no intervening sequence between the repeated elements.

Interrupted repeat:

Repeats in which there is additional sequence separating the repeated elements.

Tandem repeat:

A direct repeat in which no additional sequence separates the repeated elements.
It is therefore a continuous, direct repeat region as defined using the above terms.

Purine/pyrimidine stretch:

A region in which one strand is composed of purines or pyrimidines

Characteristics of repeat regions:

a. Motif: the composition of nucleotides making up the basic unit of the repeat region.

b. Motif length: the length of the repeated element.

for oligonucleotide tracts this includes: homopolymeric tracts, di-, tri-, tetra-, penta- etc nucleotide repeats.

c. Copy number: the number of motifs constituting the repeat region.

d. Motif prevalence: the frequency of motifs in the genome.

e. Repeat prevalence: the frequency of the repeat region in the genome.

f. Fidelity: the % identity of the repeated elements.

g. Direction on the chromosome relative to the origin of replication.

h. Coding or non-coding: whether or not the repeat is within or without an open reading frame.

i. Active: a repeat for which it has been demonstrated that a change in length results directly in an altered level of expression of a gene or the amino acid composition of the gene product with which the repeat is associated.

j. Potentially active: when the above is considered likely but for which formal proof is yet to be obtained.

k. Epitope repeat: a repeat element present within an open reading frame that does not alter the reading frame of the subsequent sequence regardless of the number of repeats present.

l. Organisational repeat: a repeat involved in genome rearrangement.